FIVE-MODALITY SEMANTIC SEGMENTATION:

Dataset, Encoder & Decoder Fusion, and Per-Pixel Gating

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

at Massey University, Auckland,

New Zealand.

Martin Brenner

2026

Contents

Abstract x

Author's Declaration xi

Acknowledgements xii

Funding xiii

Abbreviations and Terms xiv

1 Introduction 1

1.1 Introduction . . . 1
1.2 Objectives . . . 2
1.3 Research Questions . . . 2
1.4 Scope of the Research . . . 3
1.5 Thesis Overview . . . 3
1.6 Publications & Contributions . . . 4

2 RGB-D-T Fusion: A Systematic Literature Review 7

2.1 Abstract . . . 9
2.2 Introduction . . . 9
2.3 Methodology And Research Description . . . 11
2.4 Background . . . 14
2.5 Camera Calibration And Registration . . . 16
2.5.1 Calibration Boards . . . 16
2.5.2 Registration . . . 18
2.6 Thermal Data Visualisation . . . 20
2.7 How And What To Fuse . . . 20
2.7.1 Fusion Stages . . . 20
2.7.2 Fusion Methodologies . . . 21
2.8 ROI & Overlay . . . 22
2.9 Process Support . . . 24
2.10 3D Reconstruction . . . 24
2.11 Manual Descriptor-based Methods . . . 25
2.11.1 Visual Modality (V) . . . 26
2.11.2 Depth Modality (D) . . . 26
2.11.3 Thermal Modality (T) . . . 27
2.11.4 General Feature Extraction . . . 27
2.11.5 Segmentation . . . 28
2.11.6 Fusion & Evaluation . . . 29
2.12 Deep Learning-based Methods . . . 31
2.12.1 Disparity Prediction . . . 31

2.12.2 Fusion Methods (DL) . . . 32
2.12.3 Fusion . . . 34
2.12.4 Semantic Segmentation . . . 35
2.12.5 Object Detection . . . 37
2.12.6 Presentation Attack Detection (PAD) . . . 38
2.13 Datasets . . . 40
2.14 Limitations . . . 40
2.14.1 Sensors . . . 40
2.14.2 Modalities . . . 40
2.15 Synthesis . . . 42
2.16 Challenges and future work . . . 42
2.16.1 Data Fusion . . . 42
2.16.2 Thermal Data . . . 43
2.16.3 Depth Data . . . 44
2.16.4 Datasets . . . 44
2.16.5 Deep Learning . . . 44
2.16.6 PAD . . . 44
2.17 Conclusion . . . 45

3 MM5: Dataset, Capture, Calibration and Preprocessing 55
3.1 Abstract . . . 57
3.2 Introduction . . . 57
3.2.1 Key Contributions . . . 58
3.3 Related Work . . . 59
3.3.1 Existing Multimodal Datasets . . . 59
3.3.2 Multi-Resolution Thermal Encoding . . . 60
3.3.3 Multi-Resolution Depth Encoding . . . 62
3.4 Multimodal Hardware System . . . 62
3.5 Capturing Setup . . . 63
3.5.1 Lighting . . . 63
3.6 Capturing Software . . . 64
3.7 Camera Calibration And Registration . . . 64
3.8 Labelling and Post-Processing Pipeline . . . 65
3.8.1 MAR: Multimodal Annotation Remapping Algorithm . . . 66
3.8.2 Final Image Generation and Storage . . . 68
3.9 MM5 Dataset . . . 68
3.9.1 Structure . . . 68
3.9.2 Images . . . 69
3.9.3 Thermal Raw Data . . . 69
3.9.4 Thermal 8-bit Data . . . 70
3.9.5 Thermal 24-bit Data . . . 70
3.9.6 Depth Raw Data . . . 70
3.9.7 Depth 8-bit Data . . . 70
3.9.8 Depth Focused 8-bit Data . . . 70

3.9.9 Depth Focused 24-bit Data . . . 70
3.9.10 Meta Data . . . 70
3.9.11 Labels . . . 71
3.9.12 Classes . . . 71
3.9.13 Calibration Data . . . 72
3.10 Data Pre-Processing . . . 72
3.10.1 Processing Thermal Data With DTMRE . . . 72
3.10.2 Processing Depth Data With ADMRE . . . 74
3.11 Challenges and Future Work . . . 76
3.12 Conclusion . . . 77
3.A Labelling Process . . . 79
3.A.1 Labelling Process Overview . . . 79
3.A.2 Labelling and Post-Processing: Label Studio . . . 80
3.A.3 Multimodal Image Alignment and Processing . . . 80
3.B DTMRE Implementation . . . 81
3.B.1 Detailed Evaluation Results . . . 81
3.B.2 DTMRE Algorithm . . . 82
3.C ADMRE Implementation . . . 83
3.C.1 Processing Steps . . . 83
3.C.2 Detailed Evaluation Results . . . 84
3.C.3 ADMRE Algorithm . . . 85
3.D MAR Implementation . . . 86
3.D.1 Inverse Mapping . . . 86
3.D.2 Re-distortion . . . 86
3.D.3 Depth Correction . . . 87
3.D.4 Edge-Guided Random Walker Refinement . . . 87
3.D.5 Flowchart . . . 89
3.D.6 MAR Algorithm . . . 90

4 Encoder-Level Fusion for Semantic Segmentation 96
4.1 Abstract . . . 98
4.2 Introduction . . . 98
4.2.1 Key Contributions . . . 99
4.3 Related Work . . . 99
4.4 MM5 Dataset . . . 100
4.4.1 Training and Evaluation Data . . . 101
4.5 Proposed Method . . . 101
4.5.1 Encoder: Hierarchical MiT backbone with per-pixel gated multimodal fusion . . . 103
4.5.2 Data-level Fusion . . . 105
4.5.3 Stage-Wise Intensity Fusion(SWIF) . . . 105
4.5.4 Auxiliary Modality Alignment . . . 106
4.5.5 Auxiliary Modality Fusion . . . 107
4.5.6 Decoder Architecture . . . 108
4.5.7 Training Procedure . . . 109

4.6 Results and Discussion . . . 111
4.6.1 Evaluation Metrics and Comparative Analysis . . . 111
4.6.2 Overall Performance . . . 111
4.6.3 Impact of Lighting Conditions . . . 112
4.6.4 Challenging Classes: Contribution of Thermal and UV Streams . . . 114
4.6.5 Computational Requirements and Throughput . . . 114
4.6.6 Comparative Analysis of Fusion Strategies . . . 114
4.7 Failure Case Analysis and Modality Importance . . . 118
4.7.1 Failure Case Analysis with Full Multimodal Input . . . 118
4.7.2 Modality ablation study . . . 122
4.7.3 Comparative Analysis of Failure Modes . . . 124
4.8 Conclusion . . . 124
4.9 Future Work . . . 125
4.A Detailed Network Results . . . 126
4.A.1 MiT-B0 500 Epochs . . . 126
4.A.2 MiT-B2 250 Epochs . . . 127
4.A.3 MiT-B0 Comparison . . . 128
4.B Implementation Details . . . 129
4.B.1 DIN Preprocessing Details . . . 129
4.B.2 Noise Type Details . . . 129
4.C Ablation Details . . . 130

5 Decoder-Level Fusion with Dedicated Thermal/UV Heads 133
5.1 Abstract . . . 135
5.2 Introduction . . . 135
5.2.1 Key Contributions . . . 136
5.3 Related Work . . . 137
5.3.1 Alignment Handling in Decoder Fusion . . . 138
5.3.2 Uncertainty and Dynamic Reliability in Decoder Fusion . . . 138
5.3.3 Positioning of CMAG . . . 138
5.4 MM5 Dataset . . . 138
5.4.1 Training and Evaluation Data . . . 140
5.5 Proposed Methods . . . 140
5.5.1 Architecture Overview . . . 141
5.5.2 Decoder and Pre-logit Assembly . . . 141
5.5.3 Pre-logit CMAG Fusion . . . 142
5.5.4 Multi-Head Supervision . . . 144
5.5.5 MWPA (Modality-Wise Parallel Attention) . . . 144
5.6 Experimental Setup . . . 145
5.6.1 Comparison Methods and Unified Evaluation Framework . . . 145
5.6.2 Robustness Evaluation . . . 148
5.7 Results and Discussion . . . 150
5.7.1 Evaluation Metrics . . . 150
5.7.2 Overall Performance . . . 151

5.7.3 Normalisation Strategy: LayerNorm vs GroupNorm Trade-offs . . . 151
5.7.4 Lighting Condition Sensitivity . . . 152
5.7.5 Challenging Classes and Modality-Specific Contributions . . . 153
5.7.6 Computational Efficiency and Real-Time Performance . . . 153
5.7.7 Modality Importance: Ablation Studies . . . 154
5.7.8 Noise Robustness Analysis . . . 157
5.7.9 Comparative Analysis of Fusion Strategies . . . 160
5.7.10 Encoder- vs Decoder-Level Fusion . . . 161
5.7.11 Discussion . . . 165
5.8 Conclusion . . . 167
List of Abbreviations . . . 169
5.A Implementation Details . . . 170
5.A.1 Normalisation details . . . 170
5.B Detailed Network Results . . . 171
5.B.1 Class level results at 220 epochs . . . 171
5.B.2 Network drop ablation results . . . 172
5.B.3 Decoder-Level Robustness Comparison . . . 172

6 Conclusion and Future Work **176**
6.1 Conclusion . . . 176

List of Tables

2.1 Literature searches on Google Scholar . . . 12
2.2 Overview of reviewed studies . . . 15
2.3 Publicly available RGB-DT datasets. . . . 39
2.4 Publicly available Bi-Modal datasets used in studies. . . . 39

3.1 Comparative summary of multimodal datasets with more than two modalities . . . 61
3.2 Label Distribution Table . . . 71
3.3 Defined primary colour transitions used in our DTMRE demonstration. . . . 73
3.4 Thermal preprocessing performance comparison: DTMRE vs baseline methods . . . 74
3.5 Comparison of ADMRE processed, normalised, equalized, and raw intensity data . . . 76
3.6 Depth encoding comparison: mean IoU and processing time at VGA resolution . . . 76
3.7 Per-class IoU for thermal preprocessing methods with RGB1 dim light . . . 81
3.8 Segmentation showing start/end indices and the number of unique values. . . . 83
3.9 Per-class IoU for depth encoding methods with processing time . . . 84
3.10 MAR label accuracy compared to manual corrections . . . 88

4.1 Comparison of Multimodal Fusion Methods . . . 100
4.2 Per-class training-evaluation split for MM5 top level classes . . . 102
4.3 Detailed IoU and network statistics for modality combinations (MiT-B0/B2) . . . 112
4.4 Class-wise segmentation results for fusion architectures (MiT-B0, 500 epochs) . . . 117
4.5 Classes with persistent high misclassification rates under multimodal fusion . . . 119
4.6 Detailed IoU results for modality combinations (MiT-B0, 500 epochs) . . . 126
4.7 Detailed IoU results for modality combinations (MiT-B2, 250 epochs) . . . 127
4.8 Class-wise segmentation results for fusion architectures (detailed) . . . 128
4.9 Comprehensive Ablation Analysis Summary . . . 130

5.1 Overview of modality-specific heads and GT usage in fusion methods . . . 139
5.2 Decoder-level fusion methods for three-modality pre-logit integration . . . 148
5.3 Network comparison across lighting conditions and fusion architectures . . . 151
5.4 Normalisation method comparison under optimal lighting (RGB3) . . . 152
5.5 Drop ablation results: residual mIoU after removing each modality . . . 155
5.6 Performance degradation under spatial misalignment of auxiliary modalities . . . 156
5.7 Noise robustness summary across modalities, noise types, and lighting . . . 158
5.8 Modality-specific mean loss by network averaged over noise conditions . . . 159
5.9 Encoder vs decoder-level fusion comparison at VGA resolution . . . 161
5.10 Performance degradation under complete modality removal . . . 162
5.11 Decoder vs encoder fusion performance under noise perturbations . . . 163
5.12 Detailed class-level network comparison across lighting and fusion types . . . 171
5.13 Drop ablation results detailed by lighting condition . . . 172

List of Figures

2.1 Electromagnetic spectrum with focus on infrared bands . . . 9
2.2 Number of RGB-DT, RGB-T and RGB-D publications from 2012 to 2022 . . . 10
2.3 The overall structure of this paper. . . . 12
2.4 PRISMA flow diagram illustrating the search strategy and article selection phases . . . 13
2.5 Overview of methods and areas of application of the reviewed documents. . . . 13
2.6 Stereo calibration of RGB and thermal cameras using heated bi-material board . . . 16
2.7 RGB-DT calibration board made of a glass substrate and Alumina panel. . . . 16
2.8 Calibration boards from studies using RGB, thermal and depth modalities . . . 17
2.9 Image dehazing process using GoogLeNet and Monodepth with thermal data . . . 23
2.10 ContactDB examples from thermal images of hand-object contact . . . 25
2.11 Structure of proposed feature matching algorithm. . . . 30
2.12 Architectures of early, late and middle fusion schemes in deep learning . . . 32
2.13 Deep multimodal detection architecture using YOLO with late fusion . . . 36
2.14 CNN architecture with OCCL and BCE losses for PAD feature extraction . . . 38
2.15 Preprocessed images resulting from a rigid mask attack . . . 38
2.16 RGB modality challenges: similar appearance, small objects, low illumination . . . 41
2.17 Depth image challenges: messy background, incomplete data, small objects . . . 41
2.18 Thermal challenges from VDT-2048: crossover, dispersion, and reflection . . . 41
2.19 VDT-2048 thermal AGC colour shift demonstration with hot and cold bottles . . . 43
2.20 VDT-2048 visual-thermal overlay showing parallax and distortion . . . 43
2.21 Original versus equalised depth image from VTD-2048 dataset . . . 44

3.1 Multimodal sensor array with RGB, Depth, NIR, LWIR, and UV sensors . . . 63
3.2 Capturing setup in a controlled laboratory environment. . . . 63
3.3 Screenshot of capturing software with real-time overlay and rectified thermal . . . 64
3.4 Calibration board backside with partially applied copper plates. . . . 65
3.5 RGB and thermal calibration images with stereo calibration overlay . . . 65
3.6 RGB images with and without colourised label overlay . . . 66
3.7 Thermal images showing reprojection results with and without labels . . . 67
3.8 MAR refinement process from initial labels to Random Walker optimisation . . . 67
3.9 Set of images from the dataset. . . . 69
3.10 Distribution of unique label occurrences with subclass frequencies . . . 72
3.11 Comparison of MM5 DTMRE and VDT2048 AGC thermal representations . . . 73
3.12 Normalised, focused, and difference depth images from ADMRE processing . . . 75
3.13 (a) Zoomed-in normalised 8-bit image, and (b) zoomed-in focused 8-bit image of grapes. . . . 76
3.14 Labelling process pipeline . . . 79
3.15 Raw depth histogram with KDE peak detection . . . 83
3.16 RGB image of partially rotten green grapes. . . . 83
3.17 MAR Flowchart . . . 89

4.1 MM5 sample subset for frame 257 . . . 101

4.2 Encoder-decoder pipeline with stage-wise gated multimodal fusion . . . 103
4.3 Data-level Depth-Intensity-Normals (DIN) fusion components . . . 105
4.4 SWIF Module: Stage-Wise Intensity Fusion for RGB enhancement . . . 106
4.5 CM-FRM architecture for channel-wise and spatial-wise feature refinement . . . 107
4.6 Sigmoid Gate Module for auxiliary modality fusion . . . 108
4.7 MLP decoder schematic following SegFormer and CMX design . . . 109
4.8 Effect of normalisation on training dynamics and validation accuracy . . . 111
4.9 Example multimodal segmentation results for seven selected frames . . . 113
4.10 Class-wise IoU impact of adding T24 or U8 to multimodal fusion . . . 115
4.11 Architectural comparison: GF-Net Gated vs CMX FRM/FFM . . . 116
4.12 Mean IoU vs FPS scatter plot for compared networks . . . 118
4.13 Confusion matrix under RGB3 showing systematic misclassification patterns . . . 120
4.14 Spatial error distribution for Lemon Bad class under RGB3 conditions . . . 121
4.15 Examples of implemented noise types applied to RGB3 . . . 122
4.16 Heatmap of class-wise IoU changes under drops and corruptions . . . 123

5.1 Overview of CMAG architecture with three streams (RGB+DIN, LWIR, UV) . . . 136
5.2 MM5 sample unaligned image subset for frame 544 . . . 140
5.3 Overlap and cropping process for frame 544 of the MM5 raw data. . . . 141
5.4 SegFormer-style MLP decoder for primary stream . . . 142
5.5 Global Context Modality Attention (GCMA) mechanism . . . 143
5.6 CMAG fusion overview with multi-stage feature pyramids and GCMA . . . 144
5.7 Modality-wise Parallel Attention (MWPA) module . . . 145
5.8 PL-MMTM: Trimodal fusion through channel-wise squeeze-and-excitation . . . 146
5.9 PL-R2AU: Primary features as gating signals for auxiliary modalities . . . 147
5.10 Pre-logit sigmoid gating for auxiliary modality fusion . . . 147
5.11 Representative noise perturbations for robustness evaluation . . . 149
5.12 Spatial misalignment examples for thermal modality at 20 and 40 pixels . . . 150
5.13 Overview of drop ablation impact on modality importance . . . 154
5.14 Spatial shift sensitivity by magnitude and direction for thermal and UV . . . 156
5.15 Network performance heatmap for spatial shift robustness . . . 157
5.16 Network robustness to noise corruption across intensity levels . . . 158
5.17 Scenario-level noise sensitivity heatmap by modality and noise type . . . 159
5.18 Class-wise robustness comparison: PL-SIG vs GF-Net under perturbations . . . 164
5.19 Class-wise robustness heatmaps for decoder-level fusion architectures . . . 173

Abstract

Multimodal sensor fusion has become increasingly vital in perception systems, enabling richer scene understanding by combining complementary information from diverse sensing modalities. As autonomous systems, robotics, and computer vision applications demand greater robustness across varying environmental conditions, the need for effective fusion strategies has never been more pressing.

This thesis develops a comprehensive framework for robust multimodal perception, advancing through three key stages—evidence synthesis, dataset construction, and fusion architecture design—and is underpinned by four publications, three published in Q1 journals and one submitted.

The work begins with a systematic literature review of RGB-D-T (visual, depth, and thermal) fusion, examining existing datasets, calibration techniques, fusion approaches, and evaluation methods. This review highlights critical gaps that justify the development of a new dataset, benchmarks, and methodologies.

Building on these insights, the thesis introduces MM5, a multimodal dataset and processing pipeline combining five sensing modalities: RGB, depth, infrared intensity, thermal, and ultraviolet imaging. MM5 provides standardised capture, calibration, and preprocessing procedures, along with tools supporting data alignment and the labelling of unaligned data. The resource includes both raw and preprocessed data from the proposed depth and thermal preprocessing algorithms.

The thesis then presents two fusion strategies operating at different architectural levels to address different alignment scenarios. The first approach fuses aligned features at the encoder level, combining information from all five modalities through lightweight enhancements at each processing stage and pixel-level gating mechanisms. This yields robust baseline results while revealing the distinct contributions of each modality to overall performance.

The second approach operates at the decoder level and is designed explicitly for unaligned data. It employs separate processing heads for thermal and ultraviolet modalities, each trained with its own ground-truth labels. Crucially, this design handles unaligned and optically distorted inputs without requiring explicit preprocessing or geometric alignment, making the system more practical and resilient to modality-specific issues such as artefacts, occlusions, and missing signals.

Together, these contributions establish a five-modality benchmark and advance multimodal semantic segmentation through effective encoder-decoder fusion strategies.

Author's Declaration

This thesis has been prepared in accordance with Massey University's guidelines for a PhD thesis by publications. It comprises research outputs that have been published or submitted for publication to IEEE and Elsevier. Consequently, minor stylistic differences may occur between chapters. Where applicable, the full reference and publication status are stated at the beginning of each chapter; permissions and copyright notices are included as required. I, Martin Brenner, declare that this thesis, submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (Ph.D.) from the School of Mathematical & Computational Sciences, Massey University, is wholly my own work unless otherwise referenced. This document has not been submitted for qualifications at any other academic institution.

Acknowledgements

I would like to express my sincere gratitude to my supervisors for their invaluable guidance throughout this research journey. First and foremost, I thank Dr Napoleon Reyes for taking on the main supervisory role and providing steadfast support, insightful discussions, and collaborative brainstorming that shaped this work. I am deeply grateful to Dr Teo Susnjak for his expertise, thoughtful advice, and encouraging words that kept me motivated throughout the process. I also extend my thanks to Dr Andre Barczak for his guidance during the initial phase of my PhD studies and for the early discussions that helped establish the foundation of this research.

I am deeply grateful to Massey University for its generous support through a Doctoral Scholarship, and to the School of Natural and Computational Sciences for providing the facilities and resources that made this research possible.

Finally, I would like to thank my children, Lisa, Leon, Luis, Lara, and Luke, whose support and sacrifices made this work possible, and whose curiosity and interest in my research were a constant source of encouragement.

Funding

This research was supported by:

- Massey Doctoral Scholarship, provided by Massey University.

Abbreviations and Terms

AGC	Automatic Gain Control
AP	Average Precision
ATT	Channel and spatial dual attention
BAM	Block Attention Module
BBA	Bundle Block Adjustment
BCE	Binary Cross-Entropy
CANet	Co-Attention Network
CBAM	Convolutional Block Attention Module
CDM	Channel Dependency Module
CLAHE	Contrast-Limited Adaptive Histogram Equalisation
CMAF	Cross-Modal Attention Fusion
CMAG	Cross-Modal Attention with Gated Residuals
CM-FRM	Cross-Modal Feature Rectification Module
CMNeXt	Cross-Modal Next
CMX	Cross-Modal X
CNN	Convolutional Neural Network
COCO	Common Objects in Context (dataset)
CPS	Cross-Modal Prototype Sharing modules
CSA	Cross-Scale Aggregation
CT	Computed Tomography
CVPR	Conference on Computer Vision and Pattern Recognition
D	Depth
DGFM	Dual Gate Fusion Module
DGFNet	Dual Gate Fusion Network
DIN	Depth, Intensity, and Normals
DMAFM	Dual-Modal Attention Fusion Module
DNN	Deep Neural Network
DSM	Digital Surface Model
ECCV	European Conference on Computer Vision
EMI	Edge-based Mutual Information
ETFormer	Edge-Thermal Transformer
FCN	Fully Convolutional Network
FEM	Feature Enhancement Module
FFM	Feature Fusion Module
FFT	Fast Fourier Transform
FLAIR	Fluid-Attenuated Inversion Recovery
FOV	Field of View
FPS	Frames Per Second
FRM	Feature Rectification Module
FWIoU	Frequency Weighted Intersection over Union
GAP	Global Average Pooling
GAWFM	Global Attention-Weighted Fusion Module
GCMA	Global Context Modality Attention
GF-Net	GatedFusion-Net
GMFNet	Gated Multimodal Fusion Network
GMM	Gaussian Mixture Model
GN	Group Normalisation
GT	Ground Truth
HMM	Hidden Markov Model
HOF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradients
HON	Histogram of Oriented Normals
HRNet	High-Resolution Network
HSI	Hyperspectral Imaging
I	Infrared Intensity (NIR reflectance)
ICARCV	Int. Conf. on Control, Automation, Robotics and Vision
ICIP	IEEE Int. Conf. on Image Processing
ICMVA	Int. Conf. on Machine Vision Applications
ICP	Iterative Closest Point
ICPR	Int. Conf. on Pattern Recognition
ICRA	IEEE Int. Conf. on Robotics and Automation
IoU	Intersection over Union
IR	Infrared
IROS	IEEE/RSJ Int. Conf. on Intelligent Robots and Systems
KDE	Kernel Density Estimation
LBP	Local Binary Pattern
LED	Light-Emitting Diode
LF-DLM	Late Fusion Deep Learning Model
LiDAR	Light Detection and Ranging
LN	Layer Normalisation
LWIR	Long-Wave Infrared
MAR	Multimodal Annotation Remapping
MCAM	Multi-scale Cross Attention Module
MCCNN	Multi-Column Convolutional Neural Network
MEFNet	Modality Expert Fusion Network
MGFNet	Multi-Gated Fusion Network
mIoU	mean Intersection over Union
MiT	Mix Transformer
MLP	Multi-Layer Perceptron
MMTM	Multimodal Transfer Module
MPA	Mean Pixel Accuracy
MRI	Magnetic Resonance Imaging
MSC	Mean Shift Clustering
MSR	Multi-Scale Retinex
MUUFL	Multi-sensor Urban/Unstructured Fusion and Learning
MVS	Multi-View Stereo
MWPA	Modality-wise Parallel Attention
NIR	Near-Infrared
NLF	Non-Local Fusion
NMS	Non-Maximum Suppression
NUC	Non-Uniformity Correction
PA	Pixel Accuracy
PAD	Presentation Attack Detection
PCB	Printed Circuit Board
PET	Positron Emission Tomography
PHE	Plateau Histogram Equalisation
PICNet	Prototype-based Incremental Classification Network
PL	Pre-Logit
PL-MMTM	Pre-Logit Multimodal Transfer Module
PL-R2AU	Pre-Logit Recurrent Residual Attention U-Net
PL-SIG	Pre-Logit Sigmoid Gating
PR	Primary (RGB+DIN stream)
PSD	Power Spectral Density
PSPNet	Pyramid Scene Parsing Network
QSF-Net	Quality-aware Selective Fusion Network
R2AU	Recurrent Residual Attention U-Net
RANSAC	Random Sample Consensus
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
RGB-D	RGB-Depth
RGB-DT	RGB-Depth-Thermal
RGB-T	RGB-Thermal
ROBIO	IEEE Int. Conf. on Robotics and Biomimetics
ROI	Region of Interest
RPN	Region Proposal Network
SAM	Segment Anything Model
SAR	Synthetic Aperture Radar
SCSIFT	Shape-Constrained SIFT
SDM	Spatial Dependency Module
SE	Squeeze-and-Excitation
SGD	Stochastic Gradient Descent
SGFNet	Semantic Guidance Fusion Network
SIFT	Scale-Invariant Feature Transform
SIG	Sigmoid-Gated (residuals)
SLAM	Simultaneous Localisation and Mapping
SLR	Systematic Literature Review
SOD	Salient Object Detection
SPP	Spatial Pyramid Pooling
SSMA	Self-Supervised Model Adaptation
SVM	Support Vector Machine
SWIF	Stage-Wise Intensity Fusion
T	Thermal (long-wave infrared)
T1	T1-weighted (MRI)
T1ce	T1-weighted contrast-enhanced (MRI)
T2	T2-weighted (MRI)
T24	Thermal 24-bit (MM5 Dataset)
TCPSNet	Two-stage Cross-modal Prototype Sharing Network
TH	Thermal (auxiliary stream)

T-ICP	Trimmed Iterative Closest Point
TLS	Terrestrial Laser Scanning
TMFD	Tri-Modal Face Detection (dataset)
TMIWM	Triple-Modal Interactive Weighting Module
TPS	Thin-Plate Spline
U8	Ultraviolet 8-bit (MM5 Dataset)
UCTNet	Uncertainty-aware Cross-modal Transformer Network
UDFNet	Uncertainty-aware Dynamic Fusion Network
U-Net	U-shaped convolutional network for segmentation
UV	Ultraviolet
VHR	Very High Resolution
WMCA	Wide Multi-Channel Presentation Attack (dataset)

Chapter 1

Introduction

This chapter presents an introduction to the thesis, outlining the research objectives, guiding questions, scope of inquiry, and publications resulting from the contributions made in this study.

1.1 Introduction

The extraction and analysis of features from RGB images have become widely used processing techniques in computer vision, finding their way into a diverse array of industrial, commercial, and everyday applications. However, this technique exhibits limitations, primarily due to its confinement to the visible spectrum. The visible imaging range is notably narrower compared to other spectra, which underscores the potential benefits of exploring alternative non-visible spectral regions to overcome these restrictions.

While RGB cameras excel in tasks such as object recognition and scene understanding, their efficacy is heavily dependent on good lighting conditions. Complementary sensors address these limitations: depth sensors operate independently of visible light, thermal sensors detect heat radiation in complete darkness, and ultraviolet sensors reveal surface characteristics that are invisible to conventional cameras. The overarching aim of sensor fusion is to harmonise the strengths of each modality to mitigate their individual limitations [1].

Recent advances have progressed from early CNN-based approaches that fused separate backbones [2, 3] to transformer-based architectures, enabling more efficient and flexible fusion [4, 5]. Despite these developments, progress in fusing three or more modalities has been hindered by the lack of datasets. The VDT-2048 corpus [2], while marking a turning point for RGB-D-T fusion, is limited to 8-bit auto-gain thermal data with a pseudo-colour scheme applied. Most existing datasets remain constrained to two modalities, lack raw sensor data, or cover only limited scenarios [1].

Practical deployment faces significant challenges. In real capture systems, sensors exhibit heterogeneous resolutions, fields of view, and optical characteristics. Thermal and UV cameras typically use different optics and exhibit lens distortion, making geometric alignment non-trivial. Current fusion strategies often assume perfect registration or rely on learnt rectification [6, 4], increasing complexity and limiting scalability to larger sensor sets. Furthermore, the contribution of each modality varies across scenes and environmental conditions; yet, existing architectures rarely account for this variability.

This thesis addresses these fundamental challenges through a systematic progression from evidence synthesis to practical implementation. First, a comprehensive systematic review establishes the current state of RGB-D-T fusion, identifying critical gaps: the scarcity of publicly available tri-modal datasets providing raw sensor data and support for diverse fusion paradigms (aligned and unaligned, data-level through decision-level), the neglect of modality-specific preprocessing for depth and thermal streams, and the absence of real-time architectures for fusing three or more modalities. Second, the MM5 dataset is introduced, providing the first five-modality benchmark (RGB, depth, infrared intensity, thermal, ultraviolet) with both aligned and unaligned annotations, raw 16-bit sensor data, and systematic lighting variations. Third, an encoder-level fusion architecture demonstrates efficient integration of all five modalities through stage-wise enhancement and per-pixel gating, achieving real-time performance while establishing reproducible baselines. Finally, a decoder-level framework enables the fusion of unaligned, optically distorted inputs through cross-modal attention, eliminating the need for explicit geometric calibration. Six architecturally matched decoder-level variants are developed and systematically compared, with ablation studies contrasting encoder versus decoder-level fusion and evaluating robustness to modality dropout, spatial misalignment, and sensor noise.

Together, these contributions establish a complete framework for multimodal vision systems: from theoretical foundations through dataset resources to practical architectures that handle both ideal and challenging real-world conditions. The progression from tightly coupled encoder fusion to alignment-free decoder fusion provides researchers with validated approaches across the accuracy-robustness spectrum, advancing multimodal perception beyond current RGB-dominant paradigms.

1.2 Objectives

The primary aim of this research is to enable real-time fusion of three or more sensing modalities for object detection. The specific objectives are as follows:

- Conduct a systematic review and critical analysis of RGB-D-T fusion literature, identifying gaps in evaluation practices, modality coverage, publicly available datasets, and establishing how these gaps can be addressed to advance multimodal fusion research.
- Create a multimodal dataset that enables new benchmarking for preprocessing and fusion methods by providing raw sensor data for thermal, depth and NIR modalities, stereo imagery for RGB, depth, and NIR, UV imaging, reproducible capture tools, calibration artefacts, annotation protocols, and ground-truth annotations for both aligned and unaligned settings.
- Design a real-time transformer-based encoder fusion architecture for semantic segmentation that processes all five aligned modalities through geometric integration and per-pixel gating mechanisms to achieve robust multi-modal understanding.
- Design a decoder-level fusion framework with dedicated modality-specific heads and independent supervision that processes unaligned and optically distorted inputs directly, eliminating preprocessing requirements and improving tolerance to modality-specific artefacts.
- Establish reproducible baselines and comprehensive ablation studies across all five modalities to provide clear performance benchmarks and inform future research and practical deployment.

1.3 Research Questions

1. What does the current body of work on multimodal fusion reveal about datasets, calibration and registration practices, fusion strategies, and evaluation protocols, and which gaps hinder fair, reproducible benchmarking for and beyond RGB-D-T?
2. How should a reproducible capture and processing pipeline and dataset be designed to preserve sensor fidelity and support both aligned and unaligned experimentation, enabling fair comparison without enforcing a single registration choice?
3. How can transformer-based encoders be adapted to integrate five modalities for dense labelling with favourable accuracy-efficiency trade-offs (mIoU vs parameters/FLOPs/latency)?
4. To what extent can a pre-logit, decoder-stage fusion scheme with per-modality heads and supervision operate directly on unaligned and optically distorted inputs, thereby eliminating explicit cross-modal geometric alignment and other preprocessing, while maintaining accuracy and interpretability? How do its components (cross-attention, sigmoid gating, stage-wise aggregation) contribute to robustness against misregistration, modality drop-outs, and environmental degradation?
5. What is the marginal and context-dependent value of ultraviolet and thermal cues under varied illumination and noise, and how should per-pixel content-adaptive weighting compare with channel-wise mechanisms for robustness to misregistration and modality drop-out?
6. How does the choice of fusion stage—encoder-level versus decoder-level—affect accuracy, robustness to misregistration, modality dropout and sensor noise?
7. How can decoder-level fusion baselines be standardised to enable a fair, reproducible comparison of alignment-free multimodal fusion methods?

1.4 Scope of the Research

This research advances multimodal semantic segmentation by developing resources and fusion methods that exploit five complementary sensor streams processed as visual representations: RGB, depth, infrared intensity, thermal, and ultraviolet. Following a thesis-by-publications structure, the work comprises: (i) a focused systematic review of multimodal fusion; (ii) the MM5 dataset with a reproducible capture and processing pipeline; (iii) an encoder-level fusion baseline for five modalities; and (iv) a decoder-level approach with per-modality supervision that handles unaligned and optically distorted inputs. All methods are developed and evaluated primarily on MM5, supporting deployments with either aligned or unaligned data.

Research Boundaries

Problem setting. This work focuses on single-image semantic segmentation (dense pixel labelling). Object detection, instance segmentation, tracking, and temporal modelling are excluded.
Modalities. Five sensing modalities are considered: RGB, depth (D), near-infrared intensity (I), long-wave infrared/thermal (T), and ultraviolet (UV). Additional sensors, such as LiDAR, radar, event cameras, and hyperspectral imaging beyond UV/NIR, are outside the scope.
Dataset and pipeline. All experiments use the MM5 dataset, capture tooling, and calibration/annotation protocols. The pipeline supports both aligned and unaligned experimentation while preserving raw precision for depth, intensity, and thermal data, with adaptive and deterministic encodings provided where necessary.
UV modality constraint (analogue, low resolution). The ultraviolet channel is acquired from an analogue machine vision camera (Sony XC-EU50CE; $\sim$560 TV lines horizontal resolution). After digitisation to 720×576 pixels, UV is treated as a lower-resolution, optically distorted signal; the absence of native digital raw frames and the analogue capture chain limits fidelity (e.g., dynamic range and SNR) and thus the modality's downstream usability. The camera's spectral response is centred on a $\sim$365 nm (UV-A) peak, so the usable cues are predominantly the material reflectances around this wavelength.
Encoder-level methods. Transformer-based encoders are adapted for five-modality integration using lightweight, stage-wise geometric enhancement (from D+I and surface normals) and per-pixel gating mechanisms. Attention-heavy cross-attention and self-attention blocks are included only for comparison.
Decoder-level methods. A pre-logit fusion architecture with dedicated thermal and ultraviolet heads, each with independent supervision, processes unaligned and optically distorted inputs directly. This eliminates preprocessing overhead and removes the need for explicit cross-modal geometric alignment.
Evaluation protocol. Effectiveness is measured using standard segmentation metrics (mIoU, class-wise IoU, pixel accuracy). Efficiency is assessed through parameters, FLOPs and latency. Ablation studies emphasise robustness to misregistration, sensor noise, modality dropout, and illumination changes.
Learning setup. All models use supervised training with standard backbone initialisations.
Computing environment. Training and evaluation are conducted on a single workstation (Intel Core i7-13700F CPU; NVIDIA RTX 3090 GPU, 24 GB VRAM). No multi-GPU or distributed training is employed. Batch sizes and input resolutions are selected to fit this memory budget, and all throughput and latency reported in this thesis reflect this configuration.

1.5 Thesis Overview

This thesis addresses robust semantic segmentation across complementary visual and non-visible spectra through a systematic progression from evidence synthesis to dataset construction and fusion architecture design. Structured as a thesis by publications, the work establishes empirical foundations (datasets, preprocessing algorithms, evaluation protocols) and advances fusion strategies (encoder and decoder-level architectures with comprehensive robustness characterisation).
Chapter 2 presents a systematic literature review of RGB, depth, and thermal fusion, synthesising 47 primary studies through a PRISMA-guided methodology. The review identifies critical gaps in publicly available multimodal benchmarks, preprocessing practices for depth and thermal data, and real-time fusion capabilities—particularly the absence of transformer-based tri-modal architectures—justifying subsequent contributions.

Chapter 3 introduces MM5, a five-modality dataset integrating RGB, depth, infrared intensity, thermal, and ultraviolet imagery. The dataset captures diverse indoor scenes comprising real and replica objects—including items with visible defects and decay—under eight systematically varied RGB lighting conditions (shadows, dim lighting, overexposure) and three UV illumination settings, providing challenging scenarios for robust multimodal fusion. Novel contributions include: MAR (Multimodal Annotation Remapping) for label reprojection onto distorted auxiliary modalities; DTMRE (Deterministic Thermal Multi-Resolution Encoding) for stable 24-bit thermal representations; and ADMRE (Adaptive Depth Multi-Resolution Encoding) for adaptive depth quantisation. The dataset provides both aligned and unaligned annotation sets with preserved raw 16-bit measurements for the thermal, intensity and depth modalities.
Chapter 4 proposes GatedFusion-Net, the first transformer-based architecture that fuses all five modalities in real time. Key mechanisms include Stage-Wise Intensity Fusion (SWIF) for geometric enhancement, modality-wise normalisation for training stability, and per-pixel sigmoid gates for learnt weighted fusion. The chapter establishes five-modality segmentation baselines, quantifies the complementary value of ultraviolet cues, and compares encoder-level fusion strategies.
Chapter 5 introduces CMAG (Cross-Modal Attention with Gated Residuals) for alignment-free decoder-level fusion. The chapter develops six decoder-level variants with architectural parity, conducts extensive robustness evaluations across modality dropout, spatial misalignment, and sensor noise, and contrasts decoder versus encoder-level fusion. To support experimentation with geometrically unaligned inputs, the chapter curates an MM5 subset comprising raw lens-distorted thermal and UV imagery paired with MAR-reprojected ground-truth annotations, enabling direct evaluation of fusion methods that operate without explicit geometric calibration or spatial rectification.

1.6 Publications & Contributions

This thesis is based on four publications (P1–P4) that collectively realise 15 distinct contributions. The 15 contributions are organised as follows: one systematic review (C1), five dataset and preprocessing contributions (C2, C3, C4, C5, and C6), five encoder-level fusion contributions (C7, C8, C9, C10, and C11), and four decoder-level fusion contributions (C12, C13, C14, and C15).

P1: Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2023).
RGB-D and thermal sensor fusion: A systematic literature review.
IEEE Access, 11, 82410–82442. *(Q1)* – https://doi.org/10.1109/ACCESS.2023.3301119

- **(C1)** A systematic literature review of RGB–D–T fusion synthesising 47 primary studies to characterise traditional and deep-learning approaches, available datasets, calibration methodologies, and fusion strategies, identifying critical gaps in publicly available tri-modal benchmarks, preprocessing practices for depth and thermal data, and real-time fusion capabilities—notably the absence of transformer-based tri-modal architectures.

P2: Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2025).
MM5: Multimodal image capture and dataset generation for RGB, depth, thermal, UV, and NIR.
Information Fusion, 126, 103516. *(Q1)* – https://doi.org/10.1016/j.inffus.2025.103516

- **(C2)** The **MM5 dataset**: a five-modality benchmark with pixel-level semantic annotations capturing diverse indoor scenes with real and synthetic objects under eight RGB lighting conditions (shadows, dim lighting, overexposure) and three UV illumination settings, addressing the scarcity of publicly available datasets extending beyond RGB–D–thermal configurations.
- **(C3)** A **reproducible capture and processing pipeline** specifying hardware configuration, synchronised acquisition protocols, intrinsic/extrinsic calibration procedures, and geometric alignment workflows for standardised dataset construction and fair comparison under aligned and unaligned scenarios.
- **(C4) MAR (Multimodal Annotation Remapping)**: a label reprojection algorithm employing calibrated homographies and depth-aware geometric transformations to transfer ground-truth annotations from RGB imagery onto distorted thermal and UV images without exhaustive per-modality manual labelling.
- **(C5) DTMRE (Deterministic Thermal Multi-Resolution Encoding)**: a preprocessing algorithm providing stable 24-bit colour-mapped thermal representations with enhanced bit-depth allocation in high-variance

regions, improving thermal feature discriminability whilst maintaining compatibility with standard CNN and transformer backbones expecting three-channel input.

- **(C6) ADMRE (Adaptive Depth Multi-Resolution Encoding)**: a depth preprocessing technique applying gradient-based adaptive quantisation to allocate greater bit-depth precision in regions of strong spatial variation, enhancing geometric fidelity of depth cues for downstream fusion.

P3: Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2026).
GatedFusion-Net: Per-pixel modality weighting in a five-cue transformer for RGB-D-I-T-UV fusion.
Information Fusion, 129, 103986. *(Q1)* – https://doi.org/10.1016/j.inffus.2025.103986

- **(C7)** The first transformer-based architecture for unified RGB–depth–intensity–thermal–ultraviolet semantic segmentation, establishing real-time capability (41–74 fps at VGA resolution on RTX 3090) whilst processing five heterogeneous modalities within a single SegFormer-based model.
- **(C8)** A **dual-stage encoder architecture** comprising: (i) Stage-Wise Intensity Fusion (SWIF) injecting depth, intensity, and surface normals into RGB features at each encoder stage for geometric enhancement without parameter overhead; (ii) modality-wise normalisation using per-stream batch statistics to stabilise training and balance cross-modal influence; and (iii) lightweight per-pixel sigmoid gating enabling content-conditioned weighting without the computational cost of dense cross-attention.
- **(C9) Inaugural RGB–D–I–T–UV segmentation baselines** on MM5, achieving 88.3% mIoU under optimal illumination with maintained performance across adverse lighting conditions, and comprehensive ablations quantifying per-modality accuracy contributions.
- **(C10) First empirical evaluation of ultraviolet cues** for semantic segmentation, establishing UV's contribution through systematic ablation studies across varied lighting conditions, and characterising complementary yet situational value relative to other modalities.
- **(C11) Systematic fusion strategy comparison** contrasting early/data-level, feature-level, stage-wise enhancement, per-pixel sigmoid gating, and transformer-based cross-attention mechanisms through controlled ablations, quantifying accuracy–efficiency trade-offs and robustness to modality dropout.

P4: Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2025).
Pre-Logit Decoder Fusion for Five-Modality Segmentation with Unaligned T/UV Auxiliaries.
Submitted to the journal Information Fusion. – https://dx.doi.org/10.2139/ssrn.6023996

- **(C12) CMAG (Cross-Modal Attention and Gating)**: a decoder-level fusion module integrating geometrically unaligned thermal and UV streams into an RGB–depth–intensity–normals backbone via global cross-modal attention and sigmoid-gated residuals, achieving alignment-tolerant fusion without explicit geometric calibration, spatial warping, or feature rectification.
- **(C13)** A **family of decoder-level fusion baselines** adapting established mechanisms (MMTM channel-wise squeeze-and-excitation, R2AU recurrent attention gates, and sigmoid gating) to pre-logit integration, alongside a novel MWPA (Modality-Wise Parallel Attention) module. GCMA (Global Context Modality Attention) is additionally evaluated as a standalone component to isolate CMAG's constituent mechanisms. This unified framework enables controlled, architecturally matched comparisons under a shared MiT-B0 backbone and training protocol.
- **(C14)** An **extensive robustness characterisation** systematically evaluating all six decoder-level fusion architectures across modality dropout, spatial misalignment (20 px/40 px shifts), and sensor noise injection (four types, 14 intensity levels), with systematic comparison against encoder-level fusion to quantify the impact of fusion stage choice on accuracy–robustness trade-offs.
- **(C15) Curated unaligned MM5 subset** comprising raw lens-distorted thermal and UV imagery with MAR-reprojected ground-truth annotations, released with training code, inference scripts, and pretrained weights for CMAG and all decoder-level baselines, establishing a reproducible benchmark for alignment-free multimodal semantic segmentation.

Bibliography

[1] M. Brenner, N. H. Reyes, T. Susnjak, and A. L. C. Barczak. RGB-D and thermal sensor fusion: A systematic literature review. *IEEE Access*, 11:102667–102685, 2023.

[2] Kechen Song, Jie Wang, Yanqi Bao, Liming Huang, and Yunhui Yan. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. *IEEE/ASME Transactions on Mechatronics*, 28(3):1558–1569, 2022.

[3] Hongwei Wen, Kechen Song, Liming Huang, Han Wang, Junyi Wang, and Yunhui Yan. Hierarchical two-stage modal fusion for triple-modality salient object detection. *Measurement*, 218:113180, 2023.

[4] Jiyuan Qiu, Chen Jiang, and Haowen Wang. ETFormer: An Efficient Transformer Based on Multimodal Hybrid Fusion and Representation Learning for RGB-D-T Salient Object Detection. *IEEE Signal Processing Letters*, 31:2928–2932, 2024.

[5] Nianchang Huang, Yang Yang, Ruida Xi, Qiang Zhang, Jungong Han, and Jin Huang. Salient Object Detection From Arbitrary Modalities. *arXiv preprint arXiv:2405.03352*, 2024. Under review.

[6] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023.

Chapter 2

RGB-D-T Fusion: A Systematic Literature Review

The contents of this chapter are reproduced from the following article:
Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2023). RGB-D and thermal sensor fusion: A systematic literature review. IEEE Access, 11, 82410–82442. https://doi.org/10.1109/ACCESS.2023.3301119

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Martin Brenner
Name and title of main supervisor:	Dr Napoleon Reyes
In which chapter is the manuscript/published work?	2

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

The candidate was the main contributor of this work, and has done the literature review, experiments, and drafted the manuscript. The final draft was completed with the suggestions from the co-authors.

Please select one of the following three options:

◉ **The manuscript/published work is published or in press**

Please provide the full reference of the research output:

M. Brenner, N. H. Reyes, T. Susnjak and A. L. C. Barczak, "RGB-D and Thermal Sensor Fusion: A Systematic Literature Review," in IEEE Access, vol. 11, pp. 82410-82442, 2023, doi: 10.1109/ACCESS.2023.3301119.

○ **The manuscript is currently under review for publication**

Please provide the name of the journal:

○ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Digitally signed by Martin Brenner DN: cn=Martin Brenner, c=NZ, email=mb@fisag.co.nz Reason: I agree to specified portions of this document Location: Auckland Date: 2025.11.25 20:47:40 +13'00'	Main supervisor's signature:	Napoleon Reyes Digitally signed by Napoleon Reyes Date: 2025.12.01 17:08:38 +13'00'

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

2.1 Abstract

In the last decade, the computer vision field has seen significant progress in multimodal data fusion and learning, where multiple sensors, including depth, infrared, and visual, are used to capture the environment across diverse spectral ranges. Despite these advancements, there has been no systematic and comprehensive evaluation of fusing RGB-D and thermal modalities to date. While autonomous driving using LiDAR, radar, RGB, and other sensors has garnered substantial research interest, along with the fusion of RGB and depth modalities, the integration of thermal cameras and, specifically, the fusion of RGB-D and thermal data, has received comparatively less attention. This might be partly due to the limited number of publicly available datasets for such applications. This paper provides a comprehensive review of both, state-of-the-art and traditional methods used in fusing RGB-D and thermal camera data for various applications, such as site inspection, human tracking, fault detection, and others. The reviewed literature has been categorised into technical areas, such as 3D reconstruction, segmentation, object detection, available datasets, and other related topics. Following a brief introduction and an overview of the methodology, the study delves into calibration and registration techniques, then examines thermal visualisation and 3D reconstruction, before discussing the application of classic feature-based techniques and modern deep learning approaches. The paper concludes with a discourse on current limitations and potential future research directions. It is hoped that this survey will serve as a valuable reference for researchers looking to familiarise themselves with the latest advancements and contribute to the RGB-DT research field.

2.2 Introduction

The extraction and analysis of features from RGB images have become a widely used processing technique in computer vision, finding its way into a diverse array of industrial, commercial, and everyday applications. However, this technique exhibits limitations, primarily from its confinement to the visible spectrum. As illustrated in Figure 2.1, the visible imaging range is notably narrower compared to other spectra, which underscores the potential benefits of exploring alternative non-visible spectral regions to overcome these restrictions. The most significant constraint is

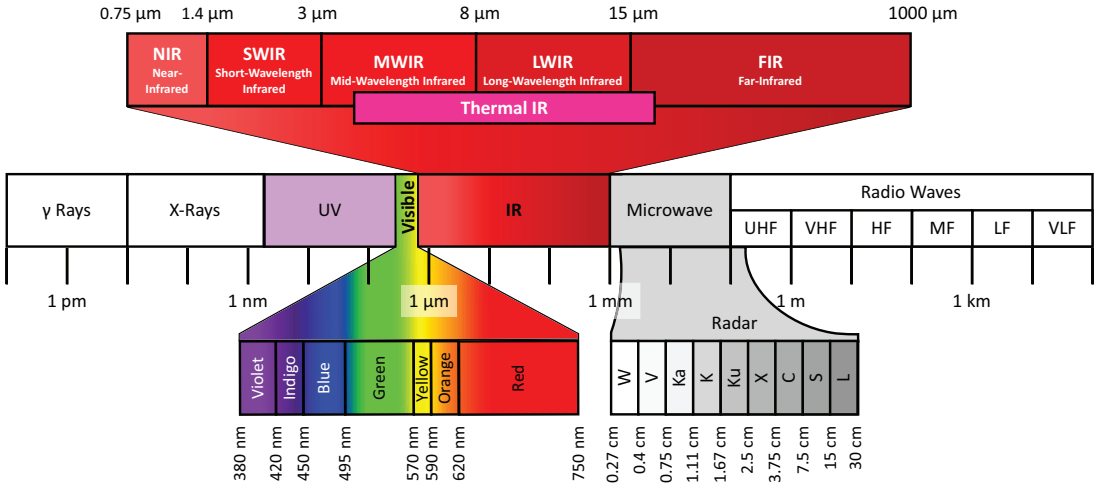


Figure 2.1: Depiction of the electromagnetic spectrum, with a focus on the various infrared (IR) bands. The thermal IR range, which is radiation-based, is specifically indicated.

that it only operates effectively under good lighting conditions and clear visibility. This has prompted researchers to explore using RGB-D and thermal cameras for multi-spectral perception in recent years as shown in Figure 2.2.
The increasing application of depth cameras can largely be attributed to the release of the Microsoft Kinect sensor in 2010. This sensor utilises an infrared (IR) structured light system, operating in the Near Infra Red (NIR) band, to capture depth information in addition to RGB colour data, and was the first depth camera to be widely available for the consumer market. Thermal cameras on the other hand capture temperature information and have been available for many years. Despite a drop in price, the cost of thermal cameras is still considerably high and the possible resolution of the sensors is low compared to RGB cameras due to the larger pixel pitch required for the Long-Wave

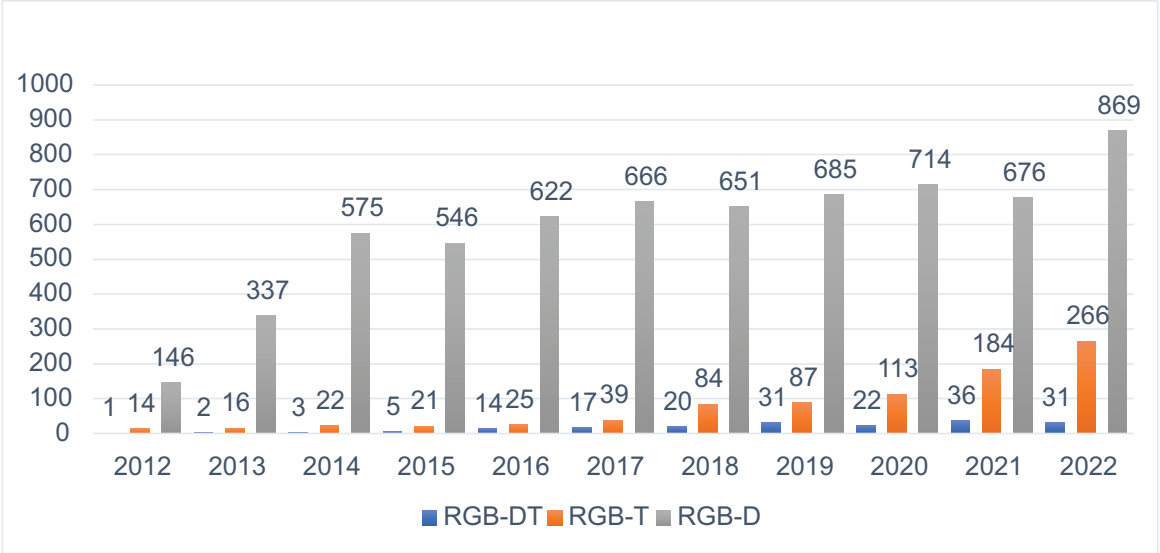


Figure 2.2: Number of publications with RGB-DT, RGB-T and RGB-D from 2012 to 2022. Source: Data from Google Scholar keyword search: ("RGB-DT" OR "RGB-DT"); ("rgb-t" OR "rgbt" OR "RGB-Thermal"); ("rgb-d" OR "rgbd" OR "RGB-Depth" OR "RGB+Depth")[peer reviewed articles only]

Infrared(LWIR) band. With the introduction of the first microbolometric array camera in 1997, detector cooling in thermal cameras became unnecessary [1], as non-cooled thermal imagers now feature electronic stabilisation. While non-cooled cameras offer advantages such as being lighter, faster, more affordable, and more reliable, cooled cameras still have the edge in terms of greater sensitivity [2].

In the context of sensor fusion, each sensor modality - RGB, Depth (D), and Thermal (T) - brings its own set of advantages and disadvantages.

RGB cameras, being ubiquitous and economical, offer high-resolution colour images that are readily interpreted by both human observers and computer vision algorithms. They excel in tasks such as object recognition, scene understanding, and texture analysis. However, their efficacy is heavily dependent on good lighting conditions.

Conversely, depth sensors, integral to RGB-D cameras, operate relatively independently of visible light, allowing them to function effectively under a variety of lighting conditions. Nevertheless, their range is typically limited, and they can be affected by factors such as sunlight interference (in the case of Time of Flight sensors) or low texture areas and lighting conditions (in the case of stereo vision). Despite these limitations, depth sensors offer valuable 3D environmental information, proving advantageous for tasks such as object detection, localisation, and navigation.

Thermal infrared sensors, in contrast to the visible and depth modalities, can sense slight temperature differences between objects and their surroundings. This capability is not hindered by low-light conditions or complete darkness, as these sensors operate based on thermal radiation, independent of any light source. This unique capability makes thermal sensing a valuable modality for object detection under challenging conditions. However, thermal sensors typically offer lower resolution than RGB cameras and are more expensive.

The overarching aim of sensor fusion is to harmonise the strengths of each sensor modality to mitigate their individual limitations. However, achieving effective fusion requires careful calibration and alignment of the sensors, along with sophisticated algorithms to integrate the different types of data.

The field of surveillance has shown significant interest in the integration of RGB and Thermal (RGB-T) data. Similarly, the combination of LiDAR sensors or stereo depth cameras with RGB, polarised images, and radar is a well-explored area in autonomous vehicles and robotics. However, the fusion of RGB-D and thermal data has not been studied as extensively in comparison. Fusion of these three modalities has the potential to provide more robust and accurate perception in various applications, such as object recognition, tracking, and localisation for applications where no long-range detection is required or the detection of endotherms is beneficial.

LiDAR and RGB-D cameras are both used for capturing 3D data, but they have different characteristics. LiDAR produces a sparser 3D point cloud with decreasing resolution over distance, while RGB-D cameras produce a more densely packed depth map that is limited to a few metres of distance. RGB-D sensors that rely on IR Time Of Flight (ToF) technology are not suitable for outdoor applications due to interference from sunlight, but devices based on

stereo vision can overcome this issue.
In order to achieve an effective fusion of the different modalities, it is essential to calibrate each sensor and align them in the same coordinate system, which involves determining the intrinsic and extrinsic parameters using the pin-hole camera model. Aligning the modalities correctly is crucial for achieving precise data fusion. Although descriptor-based methods utilising feature point matching algorithms can accomplish registration, they are often not suitable for real-time applications involving moving cameras. This is due to their high computational complexity and the challenges in implementing them with thermal data, which displays distinct characteristics compared to visual data.
Overall, the fusion of multiple modalities is an important area of research with many potential applications in various fields. With the advancements in deep learning, it is now possible to construct more advanced systems that can perform complex tasks using fused RGB-DT data.

Contribution

The primary aim of this survey is to provide a comprehensive and all-encompassing overview of the use of thermal cameras in combination with RGB and depth data. At the time of writing, the authors were unaware of comparable surveys specifically focusing on these technologies. While there are numerous reviews on sensor fusion, they predominantly focus on the amalgamation of LiDAR, Radar, RGB and other sensors, particularly within the realm of autonomous driving. These reviews often delve into the integration of these various sensor modalities and their specific challenges, yet they do not explore the specific tri-modal fusion of RGB, Depth (D), and Thermal (T) sensors. Our review uniquely situates itself at this intersection of sensor fusion, thereby distinguishing it from the broader landscape of sensor fusion literature.

The primary contributions of this review paper are designed to inform researchers working in this field by:

- Presenting a summary of various traditional and current methodologies being utilised.
- Identifying available datasets for furthering this research.
- Highlighting the current research trajectories and various application areas.

The ultimate goal is to provide a comprehensive resource that will ease the entry of interested researchers into this field while identifying trends for others.
As illustrated in Figure 2.3, the paper's structure begins with an introduction followed by a brief background to provide further context. Camera calibration and image registration are reviewed first since they are prerequisites for most approaches and fields of application. The discussion then shifts to modality fusion in general before examining the overlaying of thermal data onto visual data or 3D models for visual inspection or the extraction of thermal data from specific regions of interest. The use of one modality to support another in preprocessing is briefly addressed, followed by the exploration of RGB-DT applications in 3D reconstruction. Subsequently, the paper delves into manual descriptor-based methods and deep learning-based methods. Lastly, available datasets, limitations, and conclusions are presented.

2.3 Methodology And Research Description

The systematic literature review (SLR) for this study employed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology [3], which is a widely-used approach that involves a structured process for conducting a comprehensive literature search, applying eligibility criteria, extracting data, synthesising findings, and ensuring the search is reproducible with the same steps, keywords, and tags. The review began by defining the research topic of: RGB-D And Thermal Sensor Fusion. This was then followed by the definition of keywords and search tags used to search scientific databases via Google Scholar.

A comprehensive search resulted in the identification of 70 research papers related to the chosen topic. These papers were further refined by utilising exclusion criteria, such as language, repeated papers, and eliminating papers that were not relevant to the techniques under review, as depicted in the PRISMA flow diagram in Figure 2.4. Following the implementation of the exclusion criteria, 31 papers were reviewed in detail. Additionally, 16 more relevant documents

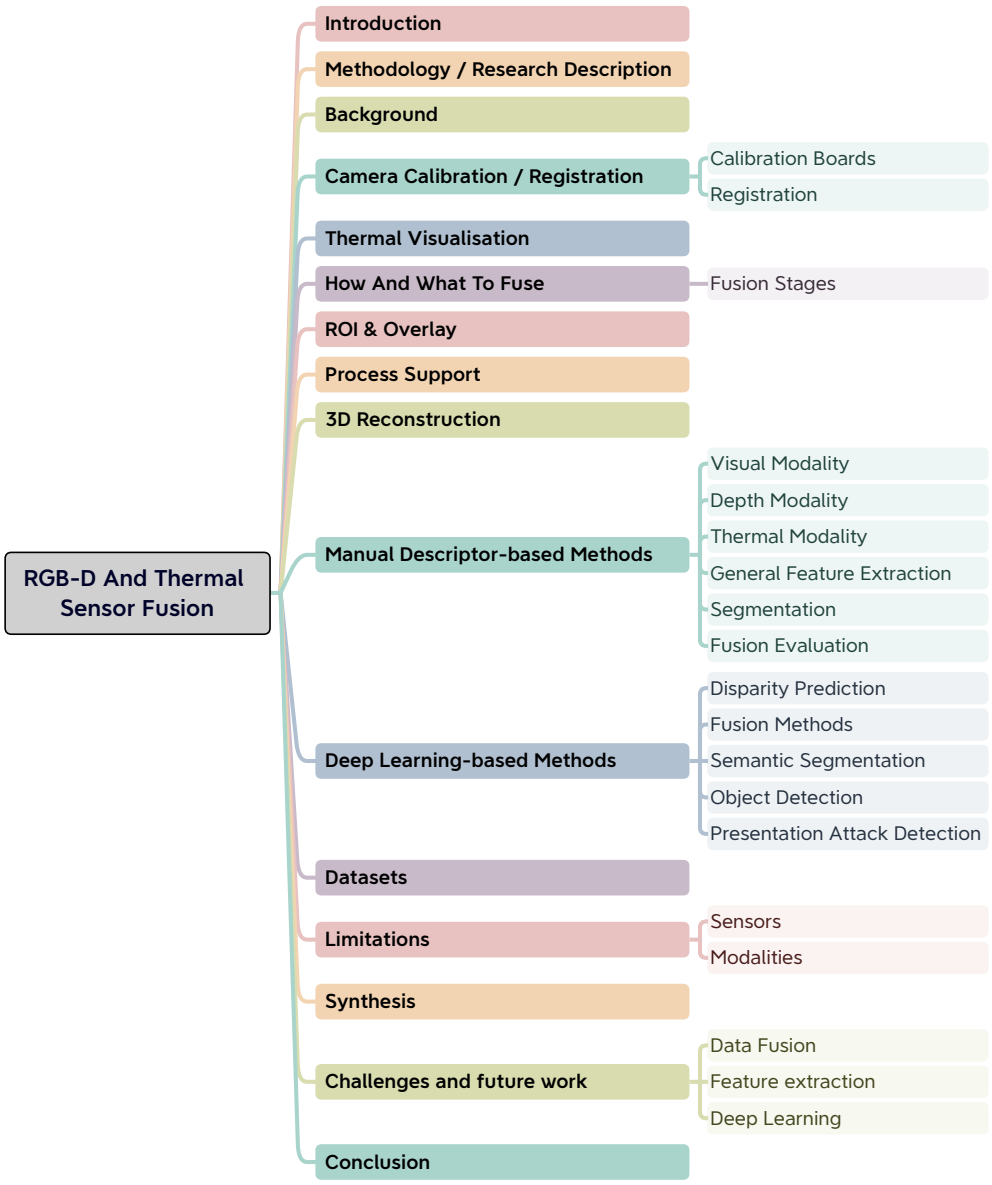


Figure 2.3: The overall structure of this paper.

Table 2.1: Literature searches on Google Scholar

Date	Terms	Filter	Results
14.1.2023	allintitle: thermal ("fuse" OR "fusing" OR "object detection" OR "object-detection" OR detection) ("3D" OR "depth" OR "point cloud" OR "point clouds")	2018	29
29.1.2023	allintitle: thermal rgb-d	-	16 (-1 duplicate)
6.3.2023	allintitle: thermal rgb "rgb-d" OR depth OR "rgb-dt" OR "rgb-d-t"	-	6(-18 duplicates)

were added after analysing the references of the initially identified papers, bringing the total number of papers reviewed to 47. Further studies that did not precisely match the three modalities, but were considered relevant to support the topic, were also included in this review. A list of the studies that have been included in the analysis can be found in Table 2.2, along with additional details such as the type of sensors used, their resolution, the frequency of data acquisition, the fusion method employed, and whether or not the system is capable of real-time processing.

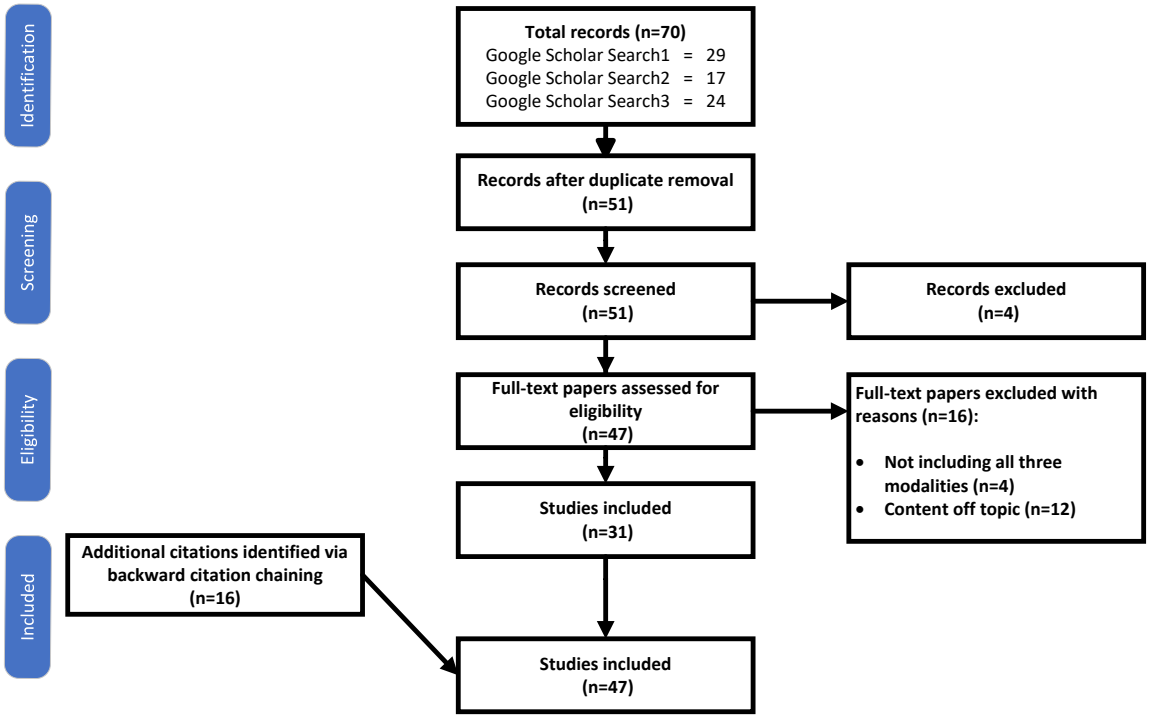


Figure 2.4: PRISMA flow diagram illustrating the search strategy and providing the phases of article identification and selection, which resulted in the identification of 47 papers that were deemed eligible for inclusion in the review. Prepared in accordance with Tricco AC, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR)[3]

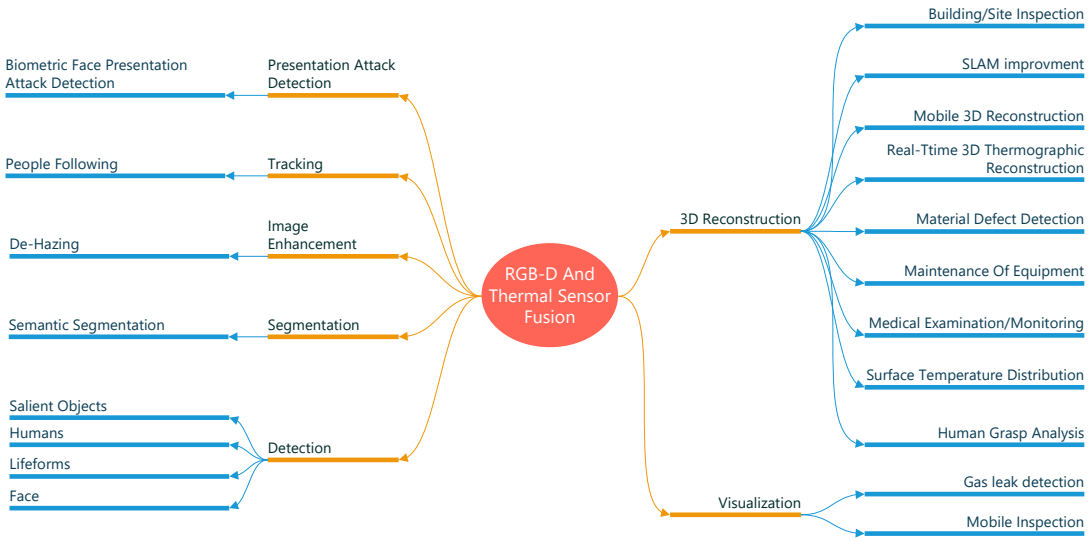


Figure 2.5: Overview of methods and areas of application of the reviewed documents.

Review questions

In this work, the aim is to answer the following review questions:

- What datasets are currently available for RGB-DT and what scenarios do they cover?
- What are the different methods for fusing the modalities?
- How are modalities weighted during fusion?
- What are the most suitable fusion and detection methods for real-time applications?
- What are the potential application areas for this technology?
- What are the limitations and future prospects?

2.4 Background

The initial research papers that concentrated on fusing RGB, Depth, and Thermal data (RGB-DT) using RGB-D cameras emerged in 2011. Early works in this field investigated medical scans [2], while later in 2013, research expanded to include 3D thermal mapping of building interiors [45], sensor fusion for people tracking [46], and tri-modal person re-identification [47].

Some earlier works proposed systems using different technologies, such as a terrestrial laser scanner and thermal infrared camera [48], or a Structure from Motion (SfM) or MultiView Stereo (MVS) pipeline to generate a dense, coloured point cloud with optional thermal data overlay [49].

Over the last decade then, the fusion of multiple modalities has been increasingly researched as combining different modalities, such as RGB-DT data, has been recognised to provide a richer and more comprehensive representation of the environment or scene. This has resulted in achieving a more accurate and robust performance in a wide range of applications, including building mapping[45], person re-identification[47], 3D salient object detection[5], autonomous driving[6, 50], activity recognition[34], robotics[5], surveillance[12], 3D reconstruction[10, 43], defect detection[29], gas leak detection [42] and many others. For example, in robotics, the combination of RGB-DT data can enable robots to perceive and navigate through complex environments with greater accuracy and efficiency [5] or to interact with humans better by interpreting their emotions[24] and activities[34]. In surveillance, the fusion of RGB-DT data has been shown to improve the detection and recognition of objects, people, and activities in a monitored area [38] under difficult light conditions, and in autonomous driving, the fusion of multi-modal data can provide a more comprehensive understanding of the surrounding environment, enabling safer and more reliable driving[50][6]. The temperature characteristics of maize under water stress, which serves as an example of multi-modal sensing in agriculture, has been investigated [11], which has the benefit of developing more efficient and sustainable agricultural practices. In the field of industrial maintenance, an Augmented Reality(AR) system, that visualises components and their temperature in real-time has been proposed [16], helping to identify faults and problems. It is also worth noting that in some cases, a single modality can indirectly improve the quality of another modality. This has been demonstrated in [6] by using the thermal data, and the extracted monodepth[51] data from the thermal data, in an algorithm used to dehaze the RGB image so that it can be used for object detection further down in the processing pipeline. Figure 2.5 offers a summary of the methods and application areas covered in the reviewed documents. This figure highlights the wide range of approaches and techniques employed in the research papers, as well as the diverse fields where these methods have been implemented.

While early works employed traditional computer vision techniques, the field has evolved alongside advancements in deep learning. Although the fusion of modalities has been shown to outperform single-modality systems, only a limited number of researchers have tackled the topic of heterogeneous sensor fusion involving stereo vision or depth cameras with thermal cameras. This is despite the growing need to meet evolving requirements and develop more robust decision-making systems by integrating features from various sensors. The potential for improved performance in a range of applications highlights the importance of continuing to explore and develop these multi-modal fusion approaches. Figure 2.2 shows the evolving trends for RGB-D, RGB-T and RGB-DT research by depicting the number of studies published over the past 10 years.

Table 2.2: Overview of reviewed studies

Ref	Year	Method	Datasets	RGB(D)	Thermal	T Res	T Hz	Type	Reg	T Use	Fus.	RT	CPU / GPU
[4]	2023	Segmentation	none	RGB Stereo	HV DS-2TD2636	384x288	50	DL	Align	DL	E	Y	Unspec/GTX 2080Ti
[5]	2022	Detection general	own public	Kinect2	FLIR A655sc	640x480	50	DL	Align	DL	MDL	N	Unspec/Unspec
[6]	2022	Image de-hazing	public	n/a	n/a			DL	Align	P/DL	L	N	Unspec/Unspec
[7]	2022	Face detection	own	Orbbec Astra	FLIR Lepton 3.5	160x120	8.7	DL	n/a	n/a	n/a	n/a	i7-6700HQ 2.6GHz/Unspec
[8]	2022	Lifeform detection	own	epc635	FLIR Lepton 3.5	160x120	8.7	F	Align	P	E	L	i7-8700/GTX 1080Ti
[9]	2022	Human detection	own	Kinect1	Seek C. Pro	320x240	8	DL	Align	DL	E	L	Unspec/Unspec
[10]	2022	3D Reconstruction	none	Realsense D455	FLIR Boson 320	320x256	60	OL	Align	P/D	E	YG	Unspec/Unspec
[11]	2021	3D Reconstruction	none	Kinect1	Optris PI400	382x288	80	SW	Feature	D	E	N	Unspec/Unspec
[12]	2021	Human detection	public	n/a	n/a			DL	Align	DL	EL	L	Unspec/2x Titan Xp
[13]	2021	Visualization	none	RGB Stereo	Seak	320x240	15	P	Align	D	OL	n/a	Unspec/Unspec
[14]	2021	3D Reconstruction	own	DJI Zenmuse XT2	FLIR XT2	640x512	9	P	Feature	D	OL	N	Unspec/Unspec
[15]	2021	Segmentation	public	n/a	n/a			DL	Align	DL	MF	Y	Unspec/Unspec
[16]	2021	3D Reconstruction	own	Realsense d415	Optris Pi640	640x480	125	DL	Align	D	OL	Y	Unspec/Unspec
[17]	2021	3D Reconstruction	none	Realsense	FLIR A65	640x512	30	ICP	Align	D	OL	YG	Unspec/Unspec
[18]	2020	3D Reconstruction	none	Photogrammetry	FLIR Zenmuse XT	640x480	30	SW	Feature	D	OL	N	Unspec/Unspec
[19]	2020	3D Reconstruction	none	Photogrammetry	FLIR A65	640x512	30	F	Feature	D	OL	N	i7-10870H/RTX 2060
[20]	2020	Tracking	none	Kinect1	FLIR Lepton 2.5	80x60	9	SW	Align	F	E	N	Unspec/Unspec
[21]	2020	Human detection	public	n/a	n/a			DL	Align	DL	EML	L	Unspec/Unspec
[22]	2020	3D Reconstruction	none	Photogrammetry	FLIR E6	160x120	9	SW	Feature	D	OL	N	Unspec/Unspec
[23]	2020	Human detection	none	Kinect1	FLIR A320	320x240	9	F	Align	D	L OL	Y	Unspec/Unspec
[24]	2020	Face detection	own	n/a	n/a			DL	Align	DL	M	n/a	Unspec/Unspec
[25]	2020	Face detection	none	Kinect1	Optris PI450	382x288	27	P	Align	F	OL	Y	Unspec/Unspec
[26]	2020	PAD	own	RealSense SR300	Seek C. Pro	320x240	15	DL	Align	DL	E M	L	Unspec/Unspec
[27]	2020	PAD	own	n/a	n/a			DL	Align	DL	E M	L	Unspec/Unspec
[28]	2019	Face detection	none	Kinect1	FLIR Lepton 2.5	80x60	9	P	Align	D	OL	Y	Unspec/Unspec
[29]	2019	3D Reconstruction	none	DAVID 3D	n/a			F	Feature	F	OL	n/a	Unspec/Unspec
[30]	2019	PAD	own	RealSense SR300	Seek C. Pro	320×240	15	DL	Align	DL	E M	L	i7-4800MQ/GTX 780M
[31]	2019	3D Reconstruction	own	Kinect2	FLIR Boson 640	640x512	30	ICP	Feature	D	OL	N	Unspec/Unspec
[32]	2018	Human detection	none	Realsense R200	FLIR Boson	n/a		P	Align	DL	L	L	i7-9700K/RTX 2060
[33]	2018	ROI Face detection	none	Asus Xtion	Optris PI640	640x480	32	SW	Align	D	OL	N	Unspec/Titan RTX
[34]	2018	Human detection	own	Kinect2	Optris PI640	640x480	32	DL	Align	F	L	N	Unspec/GTX 1080Ti
[35]	2018	3D Reconstruction	none	FLIR One	FLIR One	160x120	8.7	F	Align	D	OL	N	iPhone SE 1 (A9)/GT7600
[36]	2018	3D Reconstruction	none	Kinect2	Xenics Gobi 640	640x480	60	ICP	Align	D	OL	YG	Unspec/Unspec
[37]	2017	3D Reconstruction	own	RealSense SR300	FLIR One	160x120	8.7	P	Align	DL	E	N	Unspec/GTX 680M
[38]	2016	Tracking	none	Kinect2	FLIR A655sc	640x480	50	F	Align	F	MF	n/a	Unspec/Unspec
[39]	2016	Human detection	own public	Kinect1	AXIS Q1922	640x480	30	F	Align	F	MF	n/a	i7-960/GTX 400
[40]	2015	3D Reconstruction	none	Kinect1	Optris PI160	160x120	120	ICP	Align	D	OL	N	i7 1.9GHz/GTX 1060
[41]	2015	Face detection	none	Kinect2	AXIS Q1921	384x288	30	F	Align	D	MF	L	Unspec/Unspec
[42]	2015	Visualization	none	ASUS Xtion Pro	Optris PI450	382x288	80	F	Feature	D	OL	L	Unspec/Unspec
[43]	2014	3D Reconstruction	none	ASUS Xtion Pro	Optris PI450	382x288	80	ICP	Feature	D	OL	YG	Unspec/Unspec
[44]	2014	3D Reconstruction	none	Kinect1	Jenoptik IR-TCM	640x480	60	OL	Align	D	OL	n/a	i7-6700K 4GHz/GTX 1080
[45]	2013	3D Reconstruction	none	Kinect1	TM Miricle 307K	640x480	240	OL	Align	D	OL	L	Unspec/Unspec
[46]	2013	Tracking	none	Kinect1/Hokuyo	Heimann HTPA	32x31	9.1	F	None	F	L	Y	i7-6700K 4GHz/GTX 1080
[47]	2013	Re-identification	own	Kinect1	AXIS Q1922	640x480	30	F	Align	F	MF	L	Unspec/Titan Xp
[48]	2012	3D Reconstruction	none	Riegl VZ-400 laser	Optris PI160	160x120	120	P	Align	D	OL	n/a	Unspec/Unspec
[49]	2012	3D Reconstruction	none	RGB FLIR E60 (SfM)	FLIR E60	320x240	60	F	Feature	D	OL	n/a	Unspec/Unspec
[2]	2011	3D Reconstruction	none	Kinect1	TC384	384×288	50	OL	Align	D	OL	n/a	Unspec/Unspec

"T Use": Thermal data application; Display(D), Post-processing or Process/Algorithm(P), Feature(F), Deep Learning(DL)
"Fus.": How the Thermal data was fused with the other modalities; Late(L), Middle(M), Early(E), Overlay/Align(OL), Feature(F)
"RT": Inference speed real-time (>=30FPS); Yes(Y), Yes with GPU(YG), Likely but no data provided(L), No(N)
"Reg": Registration process: Image alignment (Align), Feature matching (Feature)
"Unspec": Unspecified

2.5 Camera Calibration And Registration

For successful multimodal environmental sensing using RGB, depth, and thermal data, it is crucial to acquire the data from these modalities in a properly aligned manner. This can pose a challenge since the sensors used for each modality may have varying fields of view (FOV), resolutions, and sensing capabilities. To facilitate data fusion, the system must be calibrated by determining the intrinsic (pin-hole camera model parameter matrix) and extrinsic (estimation of the relative sensor poses) parameters of each camera, which can then be used to align the data. This calibration, based on the pinhole camera model, has been simplified by using a stereo calibration process[52], which can be applied using these and similar modalities. This method has been implemented in numerous studies in different ways. Figure 2.6 shows the pattern matching using stereo calibration.

2.5.1 Calibration Boards

The most popular approach for the geometric calibration of thermal cameras used to be a printed chessboard heated by a flood lamp which was comparatively inaccurate and difficult to execute[53] as the temperature difference was fading quickly and the pattern was blurry. To address this a novel geometric mask with high thermal contrast that does not require a flood lamp has been proposed [53] as an alternative calibration pattern. This approach involves cutting a mask out of a thin material and holding it in front of a backdrop with a different level of thermal radiance. Building on this idea, various constructions have been developed in recent years, all based on the same principle.

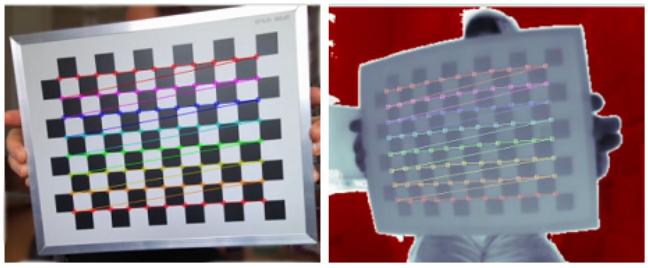


Figure 2.6: Stereo calibration of RGB and Thermal cameras. The left image shows a heated bi-material calibration checkerboard captured by an RGB camera, while the right image presents the same board as seen by a thermal camera. The overlaid lines illustrate the pattern recognition process of the stereo calibration.

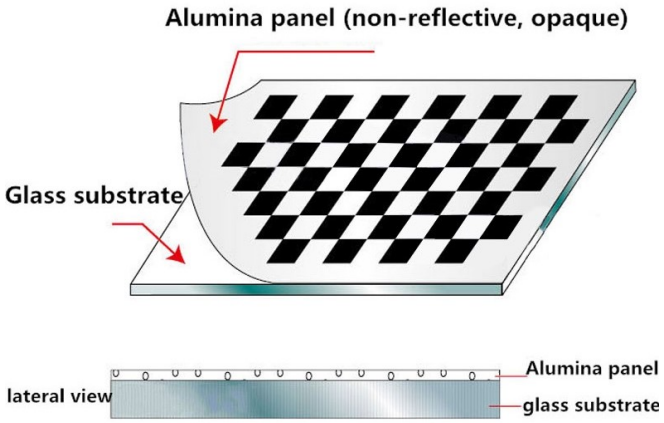


Figure 2.7: RGB-DT calibration board made of a glass substrate and Alumina panel.[5]

The multi-material calibration boards, which are essential for cross-calibrating thermal and visual modalities, with their distinct geometric patterns visible in all calibrated modalities, are used in the calibration process [52]. A checkerboard with 12 × 9 (30 mm for every square grid) with the pattern printed onto an alumina plate has been used [5] which is then mounted on a glass substrate as illustrated in Figure 2.7. The board is heated from the back, while the white reflects the heat, the black conducts it to produce the pattern in the thermal modality. These boards are commercially available. The authors in [54] constructed a board where the calibration pattern comprises a line-based grid with regularly sized square patterns. The pattern consists of thin copper lines milled onto a printed circuit board (PCB) with a width of 2 mm and a spacing of 40 mm, and it has six/seven intersections along the shorter/longer axis. Compared to conventional calibration patterns, the line-grid pattern is more robust in maintaining high contrast in thermal images due to the good conductivity of the copper lines, which ensures a uniform thermal distribution. Additionally, the proposed pattern has the same geometric relations as the conventional chessboard pattern, allowing for the use of existing algorithms for camera calibration.
However, calibration boards can be constructed simpler as demonstrated in [22] where an 11 × 11 checkerboard pattern made of cardboard paper and highly reflective metal squares was used. Alternatively, [39] constructed the calibration board using an A3-sized 10mm polystyrene foam board as a backdrop and a board of the same size with cut-out squares as the checkerboard. This is similar to [2] where a solid board was used that had rectangular holes cut out, as shown in Figure 2.8(d), whereas [13] used fabric for the black pattern. In addition to using squares, circles can also be used, as demonstrated in [8] where the authors utilised a mask made of 3mm thin Depron® material with an

asymmetric circle pattern, as shown in Figure 2.8(a), while [40] proposed using 3D printed boards in their study, as shown in Figure 2.8(g).
A distinct approach was adopted in [25] and [20], where resistors were placed onto the calibration board and heated up electrically, enabling a prolonged calibration process. Similarly, [23] and [32] employed incandescent light bulbs embedded at every other corner of the grid to emit heat.

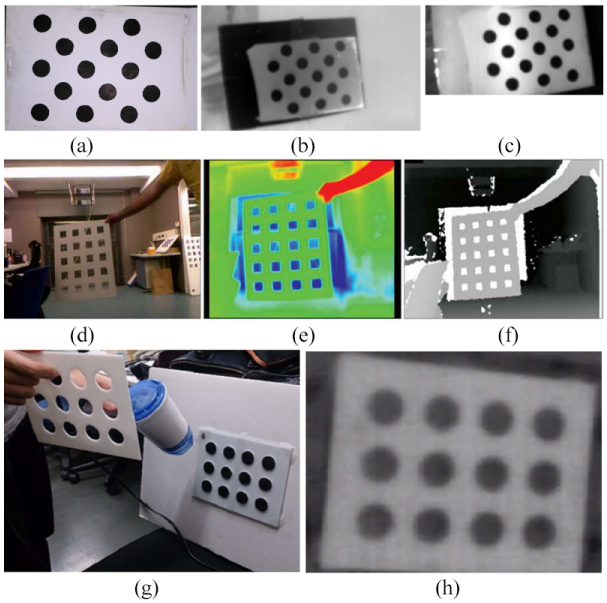


Figure 2.8: Calibration boards used in studies [8], [2] and [40] from top to bottom respectively: (a) RGB (b) Thermal (c) Depth (d) RGB (e) Thermal (f) Depth (g) Heating of lower plate (h) Thermal

In [55], a method was proposed for calibrating a UV camera with RGB-D and thermal cameras using a rectangular aluminium plate with evenly distributed circle holes. A heater strip is placed behind the plate to create sufficient contrast for the thermal camera. To ensure that all cameras can be calibrated together, a black box is used, which absorbs most of the light while allowing light to pass through the holes on the aluminium board. The white paper covering the board reflects visible and UV light, which can be detected by the RGB and UV cameras. Once the calibration tool's features are detected, the centre of each circle is marked, and OpenCV's [56] camera calibration function is used to obtain intrinsic and transformation matrices for each camera coordinate system. The proposed method allows for the accurate calibration of multiple cameras, including a UV camera, which can be beneficial in various applications. [10] used this approach to calibrate RGB-D and thermal modalities.
The combination of thermal images and colour images typically involves the use of methods that require complex calculations. However, in[57], a 2-point approach was proposed that outperformed commonly used 8-point and 7-point approaches for equalising the epipolar geometries of different images. The study proposes a method for effectively combining images by determining two points on the epipolar plane. This technique was also employed in [9] for the calibration process. In order to find calibration points in both thermal and optical data, the method used in the study involves several operations. Firstly, the Canny Edge detection method is applied to the thermal image to determine the calibration points. Next, in the optical image, the Hough circle finding method is used to locate the circles containing the calibration points, and the centres of these circles are determined as calibration points. It is important to note that the calibration mechanism design consists of two black circle drawings on a white background with incandescent bulbs at the centre of these circles. This setup allows for the creation of distinguishable common points in both the thermal and optical data, which are essential for the calibration process. Following this, line segments are extracted and plotted on both thermal and optical data. The lengths of these segments are determined by the Euclidean distance, and the slopes of the lines between the points are calculated using the slope formula and are stored for the combining process. The rotation of the thermal image is based on the difference in the calculated slopes of the lines, followed by resizing the thermal image with respect to the line length ratio. The midpoints and distances between them are obtained from thermal and optical images to achieve precise alignment in the same plane. This allows the determination of the position of the thermal image relative to the optical image[9].
In some RGB-D sensors, like the Microsoft Kinect series, the depth stream originates from a time-of-flight camera

that also generates an additional IR stream from amplitude information. Since both streams originate from the same sensor, it is referred to as the Depth/IR sensor. The IR stream can be utilised for calibration purposes, eliminating the need for any 3D elements on a board and can provide supplemental data that may be beneficial in applications for object detection or tracking in low-light conditions. This IR stream senses the 850nm (NWIR) spectral band and does not contain any thermal data. It is important to clarify that this stream should not be mistaken for the stream from a thermal camera, which is based on wavelengths of roughly $8 - 14\mu$m (LWIR). An overview of the spectral range is given in Figure 2.1.

2.5.2 Registration

RGB-D cameras, including models like Microsoft Kinect (V1, V2, and Azure) and Intel RealSense (D415, D435, etc.), are engineered to simultaneously capture both visual and depth modalities. As a result, they inherently register and output both data types. To align the thermal data, the stereo calibration process can be used to register it against the visual data.

In earlier works, before calibration using geometric patterns was applied, researchers used the Hough Parameter Space to register modalities as demonstrated in [52]. This process involved detecting edges with the Canny edge detector, resulting in binary edge images. These images were then processed by the Hough transform, which extracted all linear image segments. The rotation and translation differences could be calculated using line correspondence analysis [58]. Nonetheless, considering the two modalities as a stereo pair and employing stereo calibration techniques simplifies this process. The algorithm [52] has since been conveniently integrated into various tools such as OpenCV [56], Matlab [59], and other tools and frameworks, facilitating the acquisition of the translation vector, rotation matrix, and distortion coefficients.

The calibration of multi-camera systems, each characterised by a unique field of view (FOV), can be a challenging task, particularly when it involves a variety of modalities and resolutions. RGB and RGB-D cameras typically offer higher resolutions and distinct FOVs compared to thermal cameras. For accurate sensor fusion, optimising the overlap between the RGB, depth, and thermal modalities is crucial. In RGB-D cameras, the RGB component is usually internally pre-adjusted to match the overlapping FOV of the depth data. In sensor fusion processes that are designed for subsequent analysis and necessitate overlapping data from all modalities for real-time processing, it is necessary to modify the RGB-D data through cropping or clipping to match the resolution and FOV of the thermal camera. This requires careful consideration of the FOV of each camera during system design to ensure maximum overlap. When aligning a lower-resolution image with a smaller FOV to a higher-resolution image with a larger FOV, a homography is typically used to transform the lower-resolution image to align with the corresponding part of the higher-resolution image. This approach, which allows for the incorporation of additional information from the lower-resolution image while preserving the high-resolution data, is employed in the study [40]. In this study, the authors effectively align and fuse data from sensors with different FOVs and resolutions. They further address the challenges of occlusions and significant differences in the FOVs of the cameras, demonstrating the versatility and robustness of this approach in handling complex sensor fusion scenarios. In offline processing scenarios, more intricate techniques can be employed for alignment. For instance, the paper [14] utilises a combination of Scale-Invariant Feature Transform (SIFT) for keypoints computation and matching, Random Sample Consensus (RANSAC) for eliminating geometrically inconsistent matches, and Bundle Block Adjustment (BBA) for optimising camera parameters and producing an initial 3D structure of aligned images. Although these methods are computationally demanding, they offer superior accuracy and robustness, making them ideal for applications where precision is crucial.

While the stereo calibration approach is effective, it encounters a significant challenge from the different FOVs of the modalities, which can result in a parallax effect that varies at different depths. This phenomenon is due to the difference in viewing angles between the cameras, causing objects at various depths to appear at different positions in the different cameras. As a result, using a single homography, a transformation that maps points in one image to corresponding points in another image, only functions effectively on a specific plane. This variation in perspective leads to misalignment in the fused data.

One approach to overcoming this problem is presented in [39]. Firstly, a thermal-visible calibration device is used to establish the correspondences between the points extracted from the thermal and RGB modalities. Using a Microsoft Kinect camera, the depth sensor is already factory registered to the RGB camera; therefore, registration is focused only on the RGB to thermal data. Registration is performed using a weighted sum of multiple homographies. Multiple views of the calibration device scattered throughout the exploratory scene were used to generate homographies relating

RGB and thermal modalities. Each homography is calculated using a RANSAC-based method, taking into account the approximate distance to the view of the calibration device represented by the homography. This strategy effectively compensates for parallax at different depths. The rationale behind the approach is that registration based on each homography is only accurate for points on the plane that are spanned by the particular view of the calibration device. Therefore, to register an arbitrary point in the scene, the 8 closest homographies are weighted and then summed up. It was observed that registration accuracy is primarily dependent on 3 factors: the distance in space to the nearest homography, the synchronisation of RGB and thermal cameras, as well as the accuracy of the depth estimate.
In photogrammetry-based 3D reconstruction, as demonstrated in [14], the registration process depends on the identification and matching of keypoints, which is followed by Bundle Block Adjustment (BBA) [60]. Keypoint computation involves the detection and description of features using the SIFT [61] algorithm. Keypoints are unique locations in the image that correspond to the same real-world object across different images. The matching step entails finding matching keypoints across overlapping images. Subsequently, BBA is used to optimise the camera parameters, both internal and external, for each image, ensuring accurate calculation of ray paths inside and outside the camera for precise 3D reconstruction. These keypoint computations, matching, and BBA algorithms have been extensively studied and integrated into various software packages and frameworks for photogrammetric applications.

Automatic registration A different approach was taken by the authors in [11] by extracting edge images. To register, feature points were detected and matched. Common feature descriptors used for image registration include SIFT [61], SURF [62], and BRISK [63]. However, these methods often involve the use of a Gaussian filter, which can cause the loss of image details. To address this issue, [64] proposed a new feature descriptor called KAZE, which can detect image features in nonlinear scale spaces and obtain more feature points. The KAZE feature descriptor was utilised to register thermal and colour images of maize. The KAZE features and key points were detected from extracted edge images, and their descriptors were built. Feature points were then matched using the nearest neighbour distance ratio strategy, with outliers removed using the M-estimator Sample Consensus (MSC) algorithm, a variant of the RANSAC algorithm. This approach is akin to [29], which is elaborated in more detail in section 2.11.6. The study proposed a feature-based registration method for aligning thermal and RGB-D images using the Shape Constrained SIFT Descriptor (SCSIFT).
A similar auto registration approach was taken in [42], Edge-Based Mutual Information (EMI). However, they encountered issues when utilising the thermal images because of the Automatic Gain Control (AGC) employed in the thermal video stream. This AGC results in a variable colour range, as depicted in Figure 2.19, which shows an example from the VDT-2048 dataset. Their proposed method combines mutual information (MI), edge detection, and image separation to achieve image registration with the following steps:
Image filtering: The input images are first filtered using a Gaussian filter to reduce noise. This is done with a 9x9 kernel size and a standard deviation (σ) of 1.85.
Edge detection: A Canny edge detector is applied to both filtered images to generate edge images.
Region separation: After obtaining the edge images, region separation is performed. The primary goal of this step is to constrain the mutual information (MI) optimisation functions to focus on grey values that are in the vicinity of edges. This approach helps to ensure that the MI optimisation process is more accurate and robust, as it considers only the most relevant information in the image.
When software tools are used for the 3D reconstruction, as in [18, 22], the registration algorithms are applied by those software packages and are mainly based on a combination of feature detection, feature matching, and bundle block adjustment:
Feature detection: Identifies keypoints or features in each image. These features are typically distinct and easily recognisable patterns, such as corners, edges, or textures. The software employs SIFT (Scale-Invariant Feature Transform) or similar algorithms to extract features from the images.
Feature matching: After detecting features in each image, match corresponding features across multiple overlapping images. The software uses a matching algorithm, such as approximate nearest neighbour matching, to find the best matches between the features detected in different images.
Bundle block adjustment: Once the matching features have been identified, employ a bundle block adjustment technique to optimise the camera positions and orientations, as well as the 3D coordinates of the keypoints. This process involves minimising the reprojection error, which measures the discrepancy between the observed image coordinates and the projected coordinates of the keypoints in 3D space. Bundle block adjustment refines the initial

estimates of camera parameters and 3D points to improve the overall accuracy of the reconstructed scene.

By combining these techniques, the software registers the images, ultimately creating a consistent and accurate 3D representation of the surveyed area.

2.6 Thermal Data Visualisation

The Automatic Gain Control (AGC) technique is a histogram-based processing method that transforms raw data formats into 8-bit image data. However, this processing results in data compression, leading to a significant loss of information. In the case of 16-bit data, with a possible value range of 0 to 65,535, the resulting image is represented with values in the 0 to 255 interval, further decreasing detail. To address this issue, AGC algorithms are designed to enhance image contrast and brightness, thereby emphasising the contextual details of the scene [65].
Most LWIR cameras produce a grayscale or colour-range image stream with 8-bit per pixel. They typically use an AGC algorithm to generate the 8-bit image with high contrast. The 8-bit data represents gain-controlled values that depend on the temperature of objects in the scene and are more appropriate for human vision. However, the 8-bit representation results in a lower thermal resolution and the algorithm causes colour changes based on minimum and maximum measurements.

2.7 How And What To Fuse

In multimodal sensor fusion, deciding how and what to fuse depends on the specific application, the data modalities involved, and the desired outcome. The fusion of features or decisions can be achieved in many ways, such as concatenating feature vectors, averaging or weighted averages of data or decisions, weighted voting schemes to combine decisions, or applying machine learning techniques such as neural networks, decision trees, or support vector machines. However, the fast and massive data collection capabilities of the sensors and the representation of the obtained large data in the memory, possibly with different data types, are one of the challenges of real-time sensor fusion[9].
Alongside the fusion of different modalities, another important aspect to consider is the methodology of sensor fusion implementation. Two primary approaches dominate this field: model-based and data-driven techniques. Model-based methods, as explored in study [66], utilise pre-established models to interpret and integrate sensor data. These methods often exhibit robustness and interpretability, but their effectiveness can be constrained by the accuracy of the models they employ. In contrast, data-driven techniques, as outlined in the research [67], learn to merge sensor data directly from the data itself, typically employing machine learning techniques. These methods can potentially achieve superior performance, but they may require substantial data quantities and may be less interpretable.

2.7.1 Fusion Stages

Features of multiple modalities can be fused at different points in a process, and these fusion points are generally categorised into three levels: Data level, Feature level, and Decision level. These levels can also be referred to as low, mid, and high or early, middle, and late fusion. Each level of fusion has its advantages and disadvantages so it is essential to consider the specific context when selecting the fusion point. The three levels can be categorised as:

1. Data level(early) fusion: At the data level, the fusion of different modalities involves combining raw data from all modalities to create an integrated dataset, often by concatenating or averaging. This approach is useful when the raw data from different modalities are directly comparable and compatible. For RGB-DT data, often multi-channel images are created by blending and combining the data, primarily for deep learning purposes[9, 12, 21].

2. Feature level(middle) fusion: In this approach, features are extracted separately from each modality and then combined before being fed into a classifier or a learning algorithm. Feature-level fusion can involve concatenating the feature vectors or using other methods to merge the extracted features. This method often results in a more compact and informative representation of the data, as the features from each modality are combined after being extracted, retaining information specific to each modality. In manually crafted feature-based approaches, this is a common approach while in deep learning, this method usually enhances accuracy but has higher computational

requirements. There are many variations of middle fusion depending on the processing pipeline. The authors of [38] propose an algorithm optimised for human tracking based on an enhanced Bhattacharyya coefficient, and in [39] features are fused for body segmentation using stacked learning and Random Forest while in [41] features are combined by applying landmark-based energy filters for pain level recognition.

3. Decision level(late) fusion: At this stage, each modality is processed separately, with features extracted and then classified or analysed independently. The results or decisions from each modality are then combined to produce a final decision or output. This approach is suitable when the modalities are diverse and difficult to compare directly, or when separate classifiers have been optimised for each modality. Decision-level fusion can involve using majority voting [19, 47], weighted voting [8], or other decision-fusion techniques like a Support Vector Machine(SVM) [34].

The performance of a fusion method is highly dependent on the sensing modalities, data, and network architectures being used. This rough categorisation also holds true when applying Deep Neural Networks(DNN), which is discussed in more detail in section 2.12 and in section 2.11 for the feature-based approach. The fusion of the modalities however is not limited to a single stage but can be applied at multiple stages in a processing pipeline. Besides the listed fusion methods, it is also worth mentioning that the direct fusion of multiple modalities is not the only way to enhance the quality of data. A single modality can also indirectly enhance the quality of another modality. In [6] for example, the authors used the thermal data, together with the monodepth[51] data extracted from it, to improve the quality of RGB images by applying a dehazing algorithm.

2.7.2 Fusion Methodologies

Sensor data fusion methodologies can be broadly categorised into two main approaches: model-based and data-driven.

- **Model-Based Approaches** These methods rely on predefined models to interpret and combine sensor data. They are often robust and interpretable but may be limited by the accuracy of the models they use. Some common techniques under this category include:
  - **Kalman Filters** These are utilised in linear systems characterised by Gaussian noise, offering optimal performance in terms of minimising the mean squared error. As demonstrated in the study by the authors of [46], a Kalman filter, when combined with a probabilistic model of a leg shape, can ensure robust tracking in scenarios such as person-following.
  - **Particle Filters** These are used for non-linear and non-Gaussian systems. They are more flexible than Kalman filters but require more computational resources. In the context of person tracking, the authors of [38] employed a simple particle filter approach, which estimates the target's probability distribution using a set of weighted particles while study [20] presents an adaptive human tracking method using. The method incorporates adaptive weighting based on velocity and head position, allowing it to handle fast motion, partial occultation, and scale variation. The fusion of depth and thermal data enhances the robustness and accuracy of the tracking process, as demonstrated in various challenging scenarios.
  - **Bayesian Networks** These models are utilised in probabilistic modelling to represent the probabilistic connections between a group of variables. They are particularly beneficial when the relationships between the sensors are either known or can be learned. In the domain of Presentation Attack Detection (PAD), the authors of study [68] employed Bayesian Networks to differentiate between a genuine face and a fraudulent attack. Their approach involved designing an attack detector module based on Bayesian principles, with the decision boundary set at a log-likelihood ratio of attack to bona fide equal to 0. This design choice ensures that the classifier operates independently and maximises the confidence score in its classification.
- **Data-Driven Approaches** These methods learn to combine sensor data directly from the data itself, often using machine learning techniques. They can achieve higher performance but require large amounts of data and can be less interpretable. Some common techniques under this category include:

- **Support Vector Machines (SVMs)** These are powerful supervised learning models that perform well in high-dimensional spaces and can be customised with different Kernel functions for the decision function. However, their effectiveness can be surpassed by more complex models such as CNNs in certain contexts, as shown in [69] for fall detection systems. In the context of activity recognition and emotion classification, SVMs have demonstrated promising results when combined with various types of features. For example, the authors of [34] utilised an SVM model trained with depth and skeleton features in conjunction with thermal sensor data to enhance activity recognition accuracy. Similarly, [24] used an SVM model with both gait Power Spectral Density (PSD) and thermal features, achieving an offline testing accuracy of 70% in emotion classification while the authors of [39] compared their human body segmentation, based on Random Forest, with one using HOG + SVM. Despite the HOG + SVM approach being trained on larger, varied datasets, the study's proposed method significantly outperformed it. Further, the authors of [27], in the context of Presentation Attach Detection (PAD), noted that the SVM baseline generally performed worse than the other approaches, suggesting that the local, pixel-wise classification approach may not be as effective as the more holistic view provided by CNN models in their evaluation.
- **Decision Trees** These flowchart-like structures are used for decision-making, where each internal node signifies a test on an attribute, each branch represents the outcome of the test, and each leaf node holds a class label. They are valued for their simplicity and interpretability. For instance, the authors of [67] utilised a decision tree-based algorithm in a novel data association approach. This method used polar rays to find correspondences between trifocal camera objects and fused hypothesis, or super-sensor objects. The decision tree gradually eliminated unwanted associations by considering object characteristics such as area, visible façade, dimension ratio, and relative position in different coordinate systems.
- **Random Forest** This ensemble learning method constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. A practical application of this technique is demonstrated in [19], where a Random Forest was used to predict the conditional probabilities of different class labels based on point descriptors. The Random Forest, an ensemble of decision trees, was trained on randomly sampled subsets of training data. This approach resulted in decorrelated trees that enhanced the generalisation and robustness of the classification. The final point label was determined by majority voting across all decision trees in the Random Forest.
- **Neural Networks and Deep Learning Models** Neural networks excel at discerning complex patterns within high-dimensional data, such as images. Deep learning, a subset of neural networks, utilises multiple hidden layers to automatically learn and extract features from raw data, proving highly effective for RGB-D and Thermal sensor fusion tasks. A more comprehensive discussion on this topic can be found in section 2.12.

It's important to note that these categories are not mutually exclusive, and sensor fusion systems may use a combination of these approaches. The choice of methodology, similar to the selection of fusion methods, depends on the specific requirements of the task, the available data, and the computational resources.

2.8 ROI & Overlay

In some applications, thermal data serves as supplementary information for analysis purposes, such as site [22, 14] and building [18, 19, 43] inspections, medical examinations [23], or human thermal comfort assessments [28]. Large-area inspections for sites and buildings are generally not performed in real-time or with RGB-D sensors. Instead, photogrammetry [60] is employed, either with custom-built processing pipelines as in [14] or established tools like Pix4Dmapper, 3DF Zephyr, Context Capture, PhotoScan, and others as in [18, 22], to generate point clouds offline. By aligning thermal images, the point clouds are enriched with thermal data for offline analysis. In contrast, [35] used mobile devices and proposed image-based modelling (IBM), a passive mapping technique that uses image datasets with multiple fields of view (FOV) to reconstruct 3D models. This study employed a low-cost thermal camera and two smartphones to capture visible and thermal images. The work established that the proposed method is cost-effective and achieves a temperature precision of 2°C in the 3D thermal models, albeit at a slower pace. Since these approaches are not the primary focus of this study, they are not pursued further but are mentioned for completion as they also

represent a type of fusion of these modalities. However, the modalities are not fused to enhance a process but merely for post-analysis.

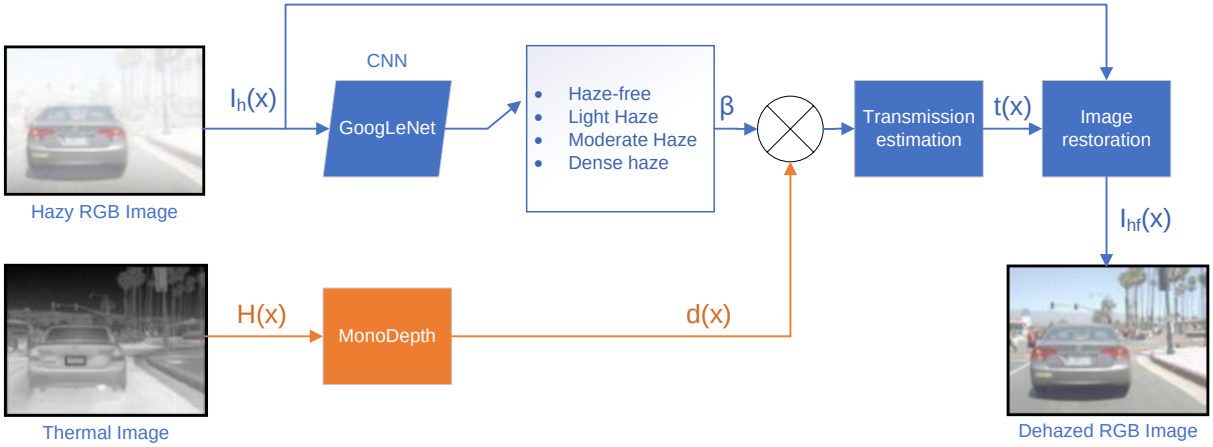


Figure 2.9: Depiction of the image dehazing process using GoogLeNet, a CNN-based classification model, to learn about weather conditions and select an appropriate atmospheric scattering coefficient based on the level of haze. The model performs depth estimation between the object and the camera using Monodepth and the thermal image. With the selected atmospheric scattering coefficient and depth information, a transmission map is estimated and a haze-free image is produced. [6]

In [4], the authors utilised stereo vision and trained a neural network for disparity estimation to generate depth data. They also applied semantic segmentation, further discussed in section 2.12.4, using depth and RGB data to define the ROI for extracting thermal data and producing a 3D reconstruction for post-processing. Meanwhile, [23] identified a region of interest(ROI) in the RGB modality also by segmentation but did this by classic methodologies not involving neural networks and extracting the thermal data by applying the ROI to the aligned thermal modality. In certain applications, such as those previously mentioned, the actual temperature values are relevant. However, in other studies like [13], the focus was on using the visual information derived from the thermal image rather than the actual temperature values. In these cases, transformations like stretching the brightness histogram values are applied to enhance the contrast, and additional denoising techniques are used to improve the image quality.
Unlike the previously discussed studies, the authors of [33] configured a system in which ROIs are identified in the RGB modality based on the facial landmark points detected using the CLM Face Tracker[70], and their coordinates are converted to thermal frame coordinates. Key regions of interest include the facial area, ocular and periocular areas, and nose area, and evaluated parameters include position, orientation, green colour component, depth (distance), and temperature. The average values of each variable are computed for each region of interest, and the relative positions and temperatures are computed with respect to the average values computed for the entire face. Finally, each computed value is logged to an individual stamped CSV file for post-experimental processing and analysis.
Similarly, in [25], face detection and extraction of landmark points from RGB images are accomplished by using the Dlib [71] machine learning toolkit based on histogram-of-oriented-gradient (HOG) features. The authors assumed that the target person does not move significantly between two consecutive frames and limited consecutive detection to the previously identified area to increase the processing speed. The facial ROIs in the thermal image are located using calibrated landmark points. The forehead centre is computed as the middle of the two eyebrow corner points, and the average temperature in the forehead area is taken as the body temperature. The mean temperatures in the nose and cheek areas are used for the measurement of the respective respiration and heartbeat rates through harmonic analysis. The dominant frequency in the temperature signal's spectrum is identified by Fast Fourier Transform (FFT), and then multiplied by 60 to obtain the respiration or heartbeat rate in cycles per minute.
For the purpose of thermal comfort of humans, the authors in [28] used algorithms implemented in OpenCV[56] for facial tracking, but unlike in [25], there was no guarantee that a face faces the camera why the thermal images used for facial skin temperature measurements contain various types of noise, such as false detection of background as faces and interference from high-temperature objects in the environment, which are represented as sudden spikes in measurements. To remove such noise, the median filter was applied before data analysis. Unlike previous studies, that segmented the frontal face into several regions and collected skin temperature from each region, this study used

global skin temperature features, including the highest, lowest, first quartile, third quartile, and average temperature measurements of all pixels in the detected facial region. These features provide an overall description of the distribution of skin temperature over a detected face, including both frontal and profile faces.
The authors in [8] adopted a different approach for processing aligned modalities. They applied background substitution and evaluated the size of connected pixel areas from the delta image to determine whether a living being was detected or not. This study fused these regions of thermal and depth data at different levels to determine the optimal result. The study did not find any significant differences in the results based on the different fusion methods used. The evaluation resulted in an accuracy of 90.1%. However, since the authors used their own data, no comparison with other methods was possible.
Numerous other studies [34, 33, 24, 41] have employed various detection methods to identify ROIs for extracting feature data to be used in decision systems or deep learning algorithms. For example, the average face temperature or the nostril area can be tracked to predict human behaviour. Further details on studies that extract data based on ROIs but process them further are presented in sections 2.11 and 2.12.

2.9 Process Support

As briefly mentioned in the Fusion Stages2.7.1 section, there is also an indirect way of using a modality to improve the quality of the data of another modality. In [6] the authors proposed a dehazing network with RGB and thermal depth (DN-RTD). To effectively remove haze, the DN-RTD dehazing network is designed to estimate β, the atmospheric scattering coefficient for the current atmospheric conditions, and $d(x)$, the depth between the camera and the object, using both RGB and thermal images. This network is shown in Figure 2.9.
In essence, the dehazing algorithm utilises GoogLeNet, a CNN-based classification model, to categorise captured hazy images $I_h(x)$ into four haze levels: haze-free, light haze, moderate haze, and dense hazy. The model then selects β that corresponds to the classified weather condition. Additionally, the algorithm estimates depth information $d(x)$ from a thermal image $H(x)$ using Monodepth, rather than an RGB image. The transmission map $t(x)$, which expresses the level of atmospheric light transmission, is derived from an Equation using the estimated β and $d(x)$. Finally, the clear image $I_{hf}(x)$ is extracted through the image restoration process. The authors then used two You Only Look Once (YOLO)[72] detectors for both, the thermal and dehazed RGB image, and fused using late fusion. However, the dehazing process takes 659.1ms to compute why it is not suitable for real-time applications yet.

2.10 3D Reconstruction

3D thermal mapping reconstruction is a crucial application area for RGB-DT images. Based on the type of 3D reconstruction equipment used, 3D thermal mapping reconstruction methods can be categorised into five groups: RGB-D (ToF or Stereo Vision), Laser Scanning, binocular stereo-structured light encoding, Photogrammetry and Structure from Motion.
The first depth camera employed to aid in 3D thermal mapping reconstruction was the Kinect v1, which has been used in various studies [40, 45, 44]. The authors in [43] developed a handheld 3D thermal mapping system using the Xtion Pro camera, and more recently, Kinect v2 and Intel RealSense[37, 10] have emerged as the most commonly used cameras for 3D thermal mapping reconstruction [36, 31, 17].
The most commonly used technique for large-scale 3D geometrical reconstruction is however Structure from Motion (SfM) [73] which was utilised in [49, 35, 48, 18]. SfM-based 3D reconstruction approaches typically extract and track robust visual features (e.g. SIFT or SURF) on 2D images captured from different viewpoints and only work well under good illumination conditions (e.g. during daytime). Feature extraction and matching, which involves the detection of SIFT features, SURF features, ORB features, and AKAZE features, is a crucial part of the SFM algorithm. However, it only produces sparse 3D point clouds, and the generated 3D models lack absolute scale information, which is not ideal for thermal diagnosis applications. To overcome these limitations, RGB-D-based 3D modelling approaches nowadays utilise depth sensors to acquire depth data of 3D objects/scenes from different viewpoints and apply 3D point cloud registration techniques, such as the iterative closest point algorithm, to align the current view with the global model [74, 75, 76, 77]. Besides the better quality, it is also worth noting that binocular stereo-structured light, as used by [78, 79], or time-of-flight depth sensors, can acquire 3D geometrical information in darkness.

Recently, the authors in [78] introduced a fast and reliable 3D thermographic reconstruction method using stereo vision. The system features adjustable measurement fields and distances, based on the chosen optics for the cameras and projector. It can reach frame rates of up to 12.5 kHz for VIS cameras and 1 kHz for the LWIR camera at full resolution. By lowering the resolution, even higher frame rates can be attained.
Meanwhile, researchers in [48] utilised terrestrial laser scanners (TLS) to acquire dense 3D point clouds, and temperature information obtained by an infrared camera is mapped onto 3D surfaces. To improve the mobility of 3D thermal imaging systems, a multi-sensor system consisting of a thermal camera and a depth sensor was built to generate 3D models with both visual and temperature information to be used for building energy efficiency monitoring [45].
Another method proposed was a thermal-guided 3D point cloud registration method (T-ICP) that improves the robustness and accuracy of 3D thermal reconstruction by integrating complementary information captured by thermal and depth sensors [36], but the method requires high computing resources to calculate several feature points. A set of experiments were performed to analyse how the key factors, such as sensing distance, specularity of the target, and scanning speed, affect the performance of high-fidelity 3D thermographic reconstruction. The authors in [10] implemented a similar idea but the localisation method combines the ORB-SLAM2 with the thermal direct method, and the entire system runs on the Robot Operation System (ROS).
Based on the Thermal-guided Iterative Closest Point (T-ICP) algorithm presented in [36], the authors of [17] developed a multi-sensor system that consists of a thermal camera, an RGB-D camera, and a digital projector. This method utilises an effective coarse-to-fine approach to enhance the robustness of pose estimation, allowing it to handle significant camera motion during large-scale thermal scanning processes. This system enables multimodal data acquisition, real-time 3D thermographic reconstruction, and spatial augmented reality through projection.
A new dataset consisting of objects and their corresponding thermal imprints resulting from grasping was proposed in [31]. To generate a coherent contact map of an object, the object is placed on a turntable which rotates as RGB-D and thermal images are captured from multiple viewpoints. The thermal images are texture-mapped onto the object's 3D mesh using a data processing technique. The steps involved in this process include extracting corresponding turntable angle and RGB, depth, and thermal images at nine locations where the turntable pauses, converting the depth maps to point clouds, estimating the turntable plane and segmenting the object using white colour segmentation, estimating the full 6D pose of the object in the nine segmented point clouds using the Iterative Closest Point (ICP) algorithm implemented in PCL, obtaining a least squares estimate of the 3D circle described by the moving object using the object origins in the nine views, and interpolating the object poses for views that are unsuitable for the ICP step. Finally, the 3D mesh along with the nine pose estimates and thermal images are input into a colourmap optimisation algorithm, which is implemented in Open3D[80], to minimise the photometric texture projection error and generate a mesh that is coherently textured with contact maps. Examples of the resulting contact maps are shown in Figure 2.10.

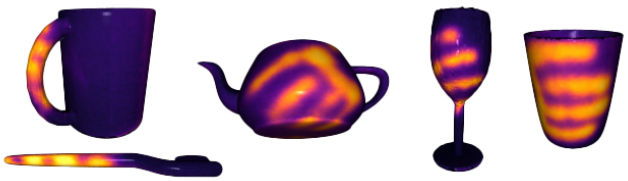


Figure 2.10: Examples from ContactDB, constructed from multiple 2D thermal images of hand-object contact resulting from human grasps.[31]

2.11 Manual Descriptor-based Methods

In contrast to deep learning methods where the feature extraction is done by the Neural Network(NN), like by convolutional layers in Convolutional Neural Networks (CNN), handcrafted descriptors are manually designed features extracted from the input data. These include histogram of oriented gradients (HOG), Histogram of Optical Flow (HOF), scale-invariant feature transform (SIFT), local binary patterns (LBP), histograms of thermal intensities, oriented gradients (HIOG), and others. These techniques are considered to be the traditional methods that have been mostly replaced by CNNs, and more recently by transformer networks, in modern detection pipelines[6]. The authors in [14] noted that the SIFT algorithm is robust and accurate for matching features in RGB images, but it only computes

low-level features and cannot recognise high-level representations.

2.11.1 Visual Modality (V)

The authors in [39] employed a combination of HOG, HOF, and HIOG to extract features. HOGs and HOFs are used to extract features from the RGB and depth data, while HIOG is used for thermal data. These features are then combined into a single feature vector and fed into a Random Forest classifier for body segmentation. The classic implementation of HOG was used for the RGB data but with a lower descriptor dimension than the original by not overlapping the HOG blocks. No gamma correction was used for the gradient computations and the Sobel kernel was applied. This means that for each pixel, the gradient orientation is determined by looking at the dominant colour channel (red, green, or blue) of that pixel, and then quantising it into a histogram over each HOG-cell [39].
HOF is a feature extraction method used to obtain motion information from an image. It works by computing dense optical flow and describing the distribution of the resultant vectors. The optical flow vectors are computed using the luminaries information of image pairs with the Gunnar Farnebäck's[81] algorithm. In [39], the authors used the implementation in OpenCV[56]. The resulting motion vectors are then masked and quantised to produce weighted votes for local motion based on their magnitude, taking into account only those motion vectors that fall inside the colour grids. The votes are locally accumulated into a v-bin histogram over each grid cell according to the signed (0°–360°) vector orientations. Unlike HOG, HOF uses signed optical flow as the orientation information provides more discriminative power [39].
Similarly [38] also used histogram-based descriptors but to process the colour modality, the RGB image is converted to a normalised colour space denoted as rgb, where $r = R/(R+G+B)$, $g = G/(R+G+B)$, and $b = B/(R+G+B)$. The colour normalisation approach is used to eliminate the illumination information in order to achieve robustness against lighting variations. Due to the fact that two components are adequate for describing the normalised colour space, with $r + g + b = 1$, a 2D histogram H_C is computed using the pair (r, g).

2.11.2 Depth Modality (D)

For depth, the authors in [39] used Histogram of Oriented Depth Normals (HON) to describe points in a point cloud. The depth modality contains a depth-dense map that represents a planar image of pixels measuring depth values in millimetres. The intrinsic parameters of the depth sensor can be used to obtain the actual coordinates from this depth representation, which can be seen as a 3D point cloud structure. This new representation allows measuring actual Euclidean distances that reflect the real world. After converting the depth modality, the surface normals for each point in the point cloud are computed, and their distribution of angles is summarised in an α-bin histogram. Then a histogram describing the distribution of the normal vectors' orientations is built. A normal vector is expressed in spherical coordinates using three parameters: the radius, the inclination θ, and the azimuth φ. In this case, the radius is a constant value, so this parameter can be omitted.
For θ and φ the calculation of the cartesian-to-spherical coordinate transformation is:

$$\theta = \arctan\left(\frac{n_z}{n_y}\right), \varphi = \arccos\frac{\sqrt{n_y^2 + n_z^2}}{n_x} \tag{2.1}$$

Thus, a 3D normal vector can be represented by a pair of angles (θ, φ), and the depth description comprises two histograms for δ_θ-bin and δ_φ-bin, which are L1-normalised and combined. These histograms describe the angular distributions of the surface normals on the body.
Similarly, [38] used an approach where a 3D normal vector is computed for each data point by fitting a 3D plane to a pre-defined local neighbourhood. Using the corresponding polar angle θ and azimuthal angle ϕ information, a 2D histogram H_D is computed.
The authors in [46] used a Leg Detection method proposed in an earlier work [82] which utilises a probabilistic leg pattern. The leg model is implemented as a sequence of maximum, minimum, maximum, minimum, and maximum values based on the laser readings, as in [83]. Various measures are defined, such as the distance between the legs and the distance between the legs and background based on these five points. Besides the laser, the depth data of an RGB-D sensor is used to detect a particular emergency vest of a person. After detecting the corners, the Lucas-Kanade

method is used to calculate the optical flow. The optical flow is computed for each corner, and in each frame, the centroid of the corners is then extracted, providing an estimation of the target's position.

2.11.3 Thermal Modality (T)

[39] used the Histogram of Thermal Intensities and Oriented Gradients (HIOG) descriptor derived from the thermal cue. This descriptor is a concatenation of two histograms. The first histogram provides a summary of thermal intensities, which are distributed over the range [0, 255]. The second histogram represents the orientations of thermal gradients. These gradients are calculated by convolving a first derivative kernel in both directions and then binned into a histogram, with their magnitude serving as a weighting factor. The two histograms are L1-normalised and concatenated. For the intensities, α_i bins, and for the gradient orientations α_g bins are used.
In a similar way, but solely relying on summarising the distribution of thermal intensities, the authors in [38] proposed a method to generate a one-dimensional histogram for the thermal modality by directly utilising the intensity values of the thermal image.
Meanwhile, [46] proposed a method to generate a 32-dimensional vector from a thermal image, where each element of the vector corresponds to the estimated probability of a person being present in a particular column of the image. This approach was chosen as the used thermal sensor had a resolution of 32x31 pixels. The computation of the vector involves three steps: firstly, a likelihood of a pixel corresponding to a person is assigned based on the assumption that the temperature of a person follows a normal distribution with mean and standard deviation values of 36 and 2, respectively, which are determined from several thermal images of people. Secondly, the likelihood matrix is smoothed by convolving it with a Gaussian kernel of a width of five pixels. Finally, the maximum value in each column of the likelihood matrix is used to determine the corresponding element of the output vector. The computation is based on established techniques such as the Lucas-Kanade optical flow method and Gaussian smoothing.

2.11.4 General Feature Extraction

The authors in [41] studied the detection of pain levels in faces and used the same feature extraction for all three modalities as a descriptor that considers both, spatial and temporal domains. This is needed to capture the spatiotemporal phenomena of changes due to pain in a facial expression. To achieve this, a steerable separable spatiotemporal filter has been selected, which utilises the second derivative of a Gaussian filter and their corresponding Hilbert transforms to measure the orientation and level of energy in the 3D space of x, y, and t. The filter provides information on the spatial texture of the face through its spatial responses and the dynamic of the features such as velocity through its temporal responses. The filter is applied independently to all three modalities, and for each pixel, the energy is calculated and normalised to improve comparability in different facial expressions. Finally, to improve localisation, the normalised energy is weighted using histograms of directions, and pixel-based energies are combined into region-based energies. For each pixel, the energy is calculated by:

$$E(x, y, t, \theta, \gamma) = [G_2(\theta, \gamma) * I(x, y, t)]^2 \tag{2.2}$$

The convolution operator '*' is used to denote the operation in which (x, y, t) represents the pixel value located at position (x, y) of the tth frame (temporal domain) in the aligned video sequence I. $E(x, y, t, \theta, \gamma)$ represents the energy released by this pixel in the direction θ and scale γ. To ensure that the obtained energy measure is comparable across different facial expressions, normalisation is performed using:

$$\hat{E}(x, y, t, \theta, \gamma) = \frac{E(x, y, t, \theta, \gamma)}{\sum E\left(x, y, t, \theta_i, \gamma\right) + \epsilon}, \tag{2.3}$$

After considering all directions θ_i, where i considers all directions and ϵ is a small bias to prevent numerical instability when the overall estimated energy is too small, the normalised energy is weighted to improve localisation using the method proposed in [84]:

$$\dot{E}(x, y, t, \theta, \gamma) = \hat{E}(x, y, t, \theta, \gamma).z(x, y, t, \theta) \tag{2.4}$$

where:

$$z(x, y, t, \theta) = \begin{cases} 1 & \sum_{\gamma_i} \hat{E}\left(x, y, t, \theta, \gamma_i\right) > Z_\theta \\ 0 & \text{Otherwise} \end{cases} \tag{2.5}$$

The resulting weighted normalised energy obtained in equation 2.4 assigns a value to each pixel based on the level of energy released by that pixel, corresponding to the chosen directions of $\theta = 0, 90, 180$, and 270. To combine these pixel-based energies into region-based energies, the authors follow study [85], by using their histograms of directions:

$$H_{R_i}(t, \theta_i, \gamma) = \sum_{R_i} \dot{E}(x, y, t, \theta_i, \gamma), \tag{2.6}$$

The histogram H_{R_i} represents the directions of the i-th region of the face, where $i = 1, 2$ or 3, and is used to combine regional histograms that are directly related to each other during the pain process. This is necessary because the muscles return to their original locations after being moved due to pain. In accordance with [85], two directions of up-down (UD) and left-right (LR) are used to combine these histograms. The directional histograms are obtained for each modality of RGB, depth, and thermal, and are subsequently used separately to determine the level of pain.

2.11.5 Segmentation

This section explores various methods of basic segmentation using multiple modalities from the reviewed studies. In [23] the authors isolate the abdominal region of newborns. The regions of interest are extracted using depth information, followed by the refinement of the human body area using the colour information to remove the background and isolate the individual. First, a dynamic depth threshold is applied to separate the body from the flat bedding surface. The distance threshold is automatically determined based on the histogram of the depth map and the first significant observed cluster according to the imaging conditions, which involve imaging the subject from above. The second step involves utilising a skin colour model that is encoded in the YCbCr space to improve the segmentation of exposed body regions from other objects in the field of view, such as probes, tubes, or clothes. The method includes multiple steps using Canny edge detection and polygonal approximation algorithms. Then, an additional refinement step is introduced in the form of a skeleton recognition method based on the depth image. This method utilises depth data to recognise various skeleton points and describe different parts of the human body.
The authors in [37] proposed a multimodal egocentric SLAM(Simultaneous Localisation and Mapping) system based on ORB-SLAM[86] which faces a significant challenge in segmenting the input frame into left-hand, right-hand, object in interaction, and static environment classes. This segmentation is crucial for two reasons: first, removing dynamic points from the input frame is essential for successful SLAM operation, and second, these labels provide the necessary structure for proper scene understanding. The semantic segmentation algorithm the authors proposed is based on priors for the hands, including their colour model, temperature, and shape. Hand location is also a prior for the object in interaction. The segmentation is performed in two steps, first segmenting the left and right hands and then the object in interaction. The right and left hands are distinguished using the prior that the right hand is at the right side of the image frame and the left hand is at the left side. CRF-based image segmentation is used to segment the hands, defining an energy minimisation problem:

$$\min_{\alpha_i^t} \sum_i U\left(\alpha_i^t, \mathbf{y}_i^t\right) + \sum_i \sum_{j \in \mathcal{N}(i)} V\left(\mathbf{y}_i^t, \mathbf{y}_j^t\right) 1\left[\alpha_i^t \neq \alpha_j^t\right] \tag{2.7}$$

In this equation, α_i^t represents a binary value of 1 if pixel i is classified as part of the hand at time t, and 0 otherwise. The neighbouring set of i is represented by $\mathcal{N}(i)$, and the indicator function is represented by $1(\cdot)$. The concatenated vector of $\mathbf{z}, \mathbf{c}, d, \tau$ is represented by $\mathbf{y}$. The unary energy function, $U\left(\alpha_i^t, \mathbf{y}_i^t\right)$, expresses the likelihood of pixel i being part of the hand, and it is a weighted combination of the probabilities of temperature (T), colour (C), hand-detector outputs (S), and history over time (H).

$$\begin{aligned} U\left(\alpha_i^t, \mathbf{y}_i^t\right) &= w_T U^T\left(\alpha_i^t, \mathbf{y}_i^t\right) + w_C U^C\left(\alpha_i^t, \mathbf{y}_i^t\right) \\ &+ w_S U^S\left(\alpha_i^t, \mathbf{y}_i^t\right) + w_H \sum_i U\left(\alpha_i^{t-1}, \mathbf{y}_i^{t-1}\right) e^{-\Delta\left(\mathbf{y}_i^t, \mathbf{y}_i^{t-1}\right)} \end{aligned} \tag{2.8}$$

where $\Delta(\cdot, \cdot)$ calculates the geodesic distance over RGB-thermal space between two voxels. $V(\cdot, \cdot)$ is a binary

consistency term that is defined over neighbouring pixels and takes the following form:

$$V\left(\mathbf{y}_i^t, \mathbf{y}_j^t\right) = \exp\left(-\frac{\left|\mathbf{y}_i^t - \mathbf{y}_j^t\right|_2}{\gamma}\right) \tag{2.9}$$

where $\gamma = \frac{1}{N}\sum_i \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}\left|\mathbf{y}_i^t - \mathbf{y}_j^t\right|_2$, and N is the total number of pixels. They defined each component of the unary energy as:

$$\begin{aligned} U^T\left(\alpha_i^t, \mathbf{y}_i^t\right) &= \tau_i^t 1\left[\alpha_i^t = 1\right] + \left(1-\tau_i^t\right)1\left[\alpha_i^t = 0\right] \\ U^C\left(\alpha_i^t, \mathbf{y}_i^t\right) &= p\left(\mathbf{c}_i^t \mid \alpha_i^t\right) \\ U^S\left(\alpha_i^t, z_i^t\right) &= \sum_{k\in\mathcal{H}} p_k e^{-\Delta\left(\mathbf{y}_i^t, \mathbf{y}_k\right)} 1\left[\alpha_i^t = 1\right] \end{aligned} \tag{2.10}$$

Where the RGB-colour model $p(\mathbf{c}_i^t|\alpha_i^t)$ is represented using a Gaussian Mixture Model (GMM) with five components and is learned separately for the hand and static scene from training data. $\mathcal{H}$ is a collection of hand detections, where each detection is represented by a centroid $\mathbf{c}_k$ and a detection likelihood $p_k \cdot \mathbf{y}_k$ which includes colour, position, depth, and temperature of the centroid of the detected hand. All components of this energy function can be computed using bi-linear filters in log-linear time and minimised using the min-cut/max-flow framework as explained in [87]. The authors used the open-source code released by the authors of [87] and the original paper provides further details. After segmenting the hands, the process continues to segment the remaining part of the image into static and dynamic object components. The same energy minimisation framework is used, with an additional motion prior and the removal of the colour prior. The motion prior accounts for the disparity between the motion of the object in interaction and the camera motion, and is defined as:

$$U^M\left(\alpha_i^t, \mathbf{y}_i^t\right) = \rho\left(\left|\mathbf{z}_i^t - \mathbf{z}^{t-1}_{\pi\left(\mathbf{R}^t\mathbf{x}_i^t + \mathbf{t}^t\right)}\right|\right) \tag{2.11}$$

In the above equation, ρ denotes the Huber function, π is the pinhole projection, $\mathbf{R}$ and $\mathbf{t}$ are the estimated camera pose, and $\mathbf{X}_i$ represents the 3D position of the i^{th} point in homogeneous coordinates. Here, α_i is a binary variable, which is equal to 1 if the i^{th} pixel belongs to the object in interaction and 0 otherwise. The tradeoff parameters $\omega_T, \omega_S, \omega_H, \omega_M$ were learned by cross validation.

2.11.6 Fusion & Evaluation

In this section, we examine the various approaches adopted for fusing the extracted features and the evaluation processes of the reviewed studies.

While [39] applied a background removal algorithm based on V and D for segmentation purposes to define the Ground Truth(GT), the authors in [20] used background removal based only on D for their body segmentation in a fixed camera set-up. The paper proposes a modality fusion method for head tracking using D and T information for fall detection. This approach uses a particle filter to estimate the head position based on both, the D and T data. A Silhouette is constructed based on D and basic body shape assumptions, while the thermal data is used to distinguish the head from the background. The fusion is performed by combining and weighting D and T based on their reliability. The authors evaluated 4 different models and concluded that the D and T data was improving the results. However, the method was limited to 8 FPS and the authors further concluded to use Deep Learning models for future refinements of this application. In [39], the authors evaluated uni-modal classification and a multi-modal fusion based on a Random Forest classifier to achieve a human body segmentation with the extracted features discussed in section 2.11.2, 2.11.1, and 2.11.3.

[38] proposed a simple person-tracking algorithm that combines the parameters of the three modalities in a way that gives less weight to modalities where camouflaging occurs. The tracker's ability to resist significant radial motions was demonstrated using the Jaccard index. The target model was modelled using a single histogram for each data source and a histogram of 3D normals was used as the depth descriptor. This did not significantly improve the tracker's accuracy compared to the same approach without depth descriptors but the authors argued it could improve its robustness in more complicated sequences.

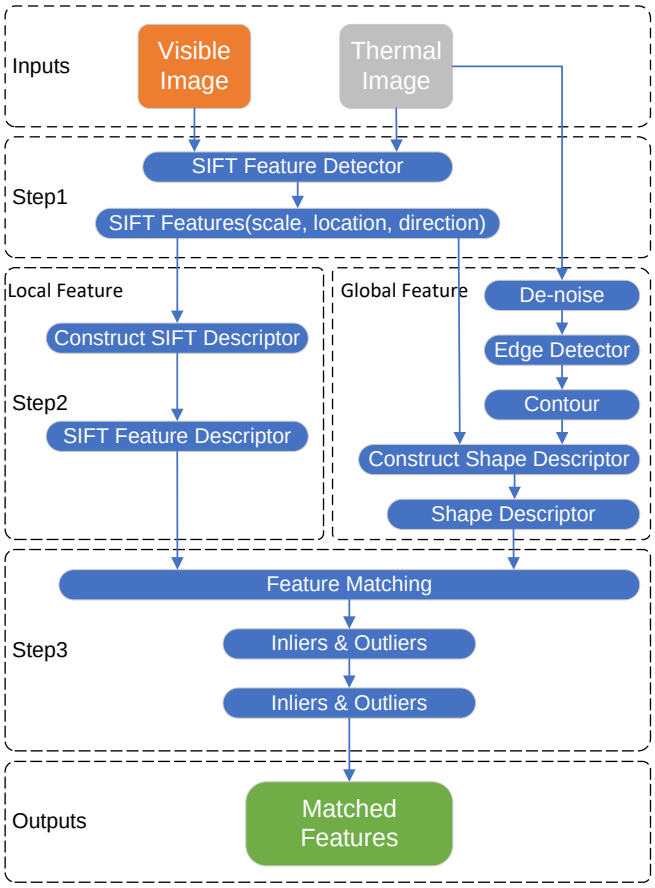


Figure 2.11: Structure of proposed feature matching algorithm.[29]

The People tracking system proposed in [46] used four modalities and is based on a laser sensor, a thermal sensor and an RGB-D camera in a mobile setup. These sensors supply input to three detection units: Leg detection, Vest Detection, and Thermal Detection, which have been discussed previously. Once their individual likelihoods are calculated, the final likelihood is calculated using coefficients to weigh the three likelihoods. The inclusion of these coefficients enables assigning more importance to one of the information sources if desired, and the authors determined these values during the evaluation process.

The researchers in [29] propose a feature-based registration method to register thermal and RGB-D images using the Shape Constrained SIFT Descriptor (SCSIFT). The registration process involves three steps: feature detection, feature description, and feature matching. In the first step, SIFT detector is applied to extract SIFT features from both visible and thermal images. In the second step, the proposed SCSIFT descriptor is constructed by combining the traditional SIFT descriptor with the shape descriptor extracted from the thermal image. In the third step, feature matching is performed by calculating the Euclidean distance between each shape descriptor vector and each SIFT descriptor vector, followed by normalisation and RANSAC to eliminate outlier matches. A detailed explanation of each this step can be found in the original study, Figure 2.11 depicts the proposed algorithm.

Shape Feature Description: Global descriptors to support local descriptors are added in multi-modality image feature matching. As the thermal image is noisy, anisotropic diffusion is applied for effective smoothing before edge extraction. The canny edge operator is used to extract edges, but contour-based methods alone are insufficient for correct feature matching. A circular template is generated around each feature point, with evenly plotted bins for edge point fitting. To describe the global position of the feature point and construct the shape descriptor, a spiral of Theodorus is applied to build the weighting function. The weighting of close region pixels is enhanced while the weighting of far region pixels is suppressed. The proposed SCSIFT descriptor is constructed from the entire image and adds a global shape constraint to the traditional SIFT descriptor which uses only local neighbourhood information.

Feature Matching Scenario Based on SCSIFT Descriptor: Normalisation is necessary before implementing RANSAC because the global shape descriptor vectors and local SIFT descriptor vectors are the statistical analysis of different information. For each feature i in the source image f_S^i with the descriptor denoted as d_S^i, the Euclidean

distance of the global descriptor $d^i_{S(G)}$ and local descriptor $d^i_{S(L)}$ to all global and local features descriptors $D^{all}_{\text{ref}\,(G)}$ and $D^{all}_{\text{ref(L)}}$ in the reference image are calculated respectively, denoted as set $\boldsymbol{E}^i_G$ and set $\boldsymbol{E}^i_L$ shown in Eq. 2.12 and 2.13.

$$\boldsymbol{E}^i_{\boldsymbol{G}} = \sqrt{\left(D^{all}_{\text{ref}\,(G)} - d^i_{S(G)}\right)^2}^2 \tag{2.12}$$

$$\boldsymbol{E}^i_L = \sqrt{\left(D^{all}_{\text{ref(L)}} - d^i_{S(l)}\right)^2}^2 \tag{2.13}$$

The ratio of the maximum value of sets $\boldsymbol{E}^i_G$ and $\boldsymbol{E}^i_L$ in Eq. 2.14 represents the scaling factor S^i. As the process of calculating Euclidean distances for descriptors is already an integral part of feature matching, this normalisation step does not add to the computational complexity.

$$S^i = \frac{\max\left(E^i_L\right)}{\max\left(E^i_G\right)} \tag{2.14}$$

And the normalisation process is done by the Eq. 2.15

$$\boldsymbol{E}^i = E^i_G \cup \frac{E^i_L}{S^i} \tag{2.15}$$

The unified distance set $\boldsymbol{E}^i$ is used to determine the most likely match for f^i_S to all features in the reference image, with the minimum value in $\boldsymbol{E}^i$ indicating the best match. Based on the maximum global and local distances, an appropriate scaling value is calculated for each feature to improve the matching accuracy. RANSAC is then employed to eliminate any outlier matches and refine the image transformation.

2.12 Deep Learning-based Methods

Deep learning-based approaches for multi-modal sensor fusion have gained increasing attention due to their ability to learn complex relationships between the different modalities and effectively fuse the information from multiple sensors. However, despite that images from multiple modalities can be beneficial in highlighting salient regions and providing more comprehensive information, they can also introduce interference between the different modalities[5]. The application of Deep Neural Networks(DNN) can be categorised into semantic segmentation and object detection. In contrast to semantic segmentation, where multi-modal features are fused at various stages within the Fully Convolutional Network (FCN), object detection involves a wider range of network architectures and fusion variants. This diversity allows for greater flexibility and adaptability in addressing specific challenges related to object detection tasks[50].
Convolutional Neural Networks (CNN) have long been the dominant architecture for image processing tasks. However, recent developments in applying transformer networks[88] to Computer Vision (CV), known as Vision Transformers (ViT)[89], have demonstrated high performance in segmentation, recognition, and detection tasks. These advances indicate the growing potential for transformer-based approaches in the field of CV.
Within the realm of multimodal object detection and segmentation, a considerable amount of research is directed towards autonomous driving applications, where the fusion of LiDAR point cloud data and RGB camera data is crucial. However, this paper focuses on RGB-D sensor data, which incorporates pre-aligned depth data or can be aligned using stereo calibration techniques. As a result, the methods for aligning point clouds with RGB data will not be discussed in this paper. However, once the RGB-D data is transformed into a point cloud, the succeeding methods for object recognition and detection can still be utilised. Although, papers using RGB stereo vision applying disparity prediction, briefly discussed in Section 2.12.1, are included as this method generates similar data as RGB-D sensors.

2.12.1 Disparity Prediction

To obtain depth data from a stereo image the disparity can either be computed, like in OpenCV which implements the block matching algorithm for calculating disparity with stereo calibration, or by training Neural Networks like

AAnet (Atrous Adaptive Network)[90]. AANet is a deep learning approach for stereo matching that can provide more accurate results than traditional methods such as block matching or simple disparity calculation. It can handle occlusions, textureless regions, and large baselines better than traditional methods. AANet is also more robust to lighting changes and can handle different camera configurations. Additionally, it can learn from large amounts of data, making it more adaptable to a wide range of scenarios. Overall, AANet provides a more flexible and accurate solution for stereo matching compared to traditional methods. Similar techniques can be applied to monocular vision to produce depth data as applied in [6] using monodepth[51].
[4] utilised AANet for disparity prediction of chicken images for feather damage analysis. The network extracts the down-sampled feature pyramid and constructs multi-scale 3D cost volumes[91]. The cost volumes are then aggregated with six stacked Adaptive Aggregation Modules (AA Modules). Each AA Module consists of three Intra-Scale Aggregation (ISA) and a Cross-Scale Aggregation (CSA). The multi-scale disparity predictions are regressed by the soft argmin mechanism and hierarchically up-sampled and refined to the original resolution. The pre-trained AANet model for the Scene Flow dataset was used for direct inference on the dataset. The dataset was augmented by random colour augmentations and vertical flipping. The initial learning rate of the pre-trained AANet model was set to 0.001 and decreased by half at 400th, 600th, 800th and 900th epochs. Adam was used to optimise the parameters of the network to minimise the average loss of the model on the training data. The disparity range was from 0 to 192 pixels.

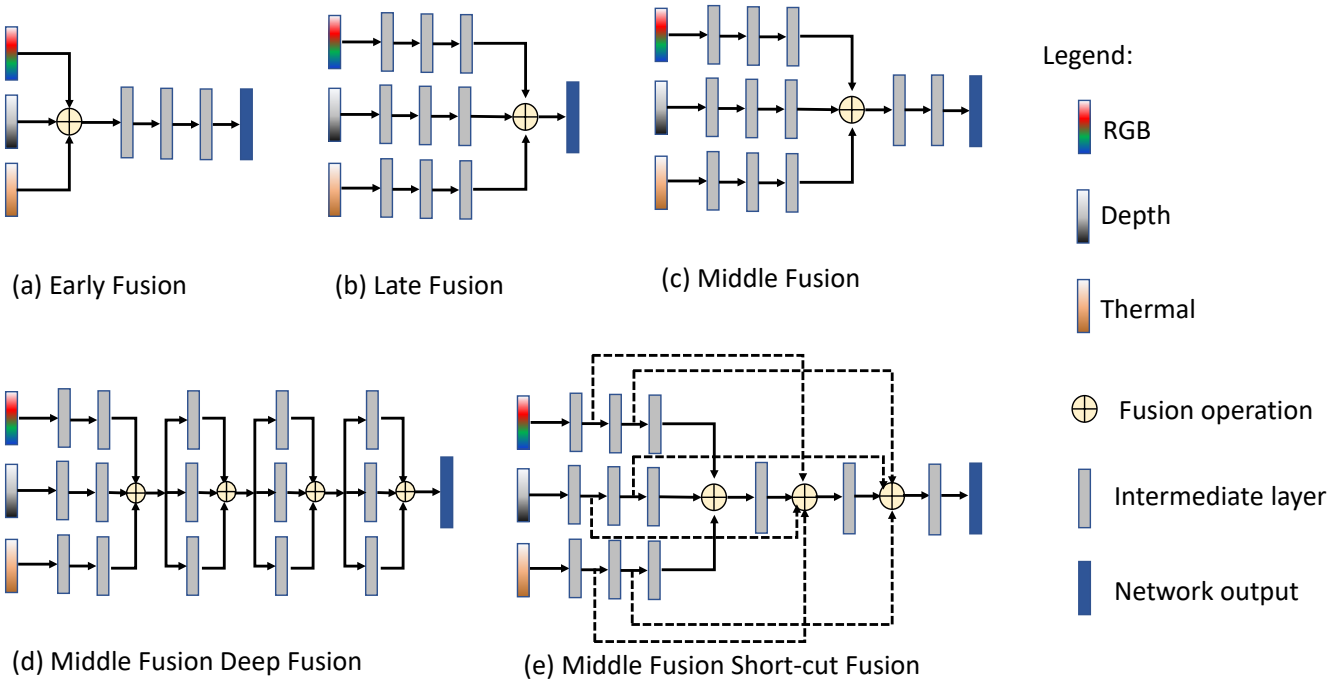


Figure 2.12: A depiction of architectures of different fusion schemes:
early fusion, late fusion, and various middle fusion techniques employed in deep learning environments based on [50].

2.12.2 Fusion Methods (DL)

Data level fusion (DL) In study [50], two primary advantages of early fusion were identified. Firstly, the network learns the joint features of multiple modalities at an early stage, allowing it to fully utilise the information present in the raw data. Secondly, early fusion has lower computational demands and requires less memory, as it processes multiple sensing modalities together. However, these benefits come at the cost of reduced model flexibility. For instance, when an input is replaced with a new sensing modality or the input channels are extended, the early fused network must be completely retrained which was also noted by authors in [12]. This study created a two-channel image out of the two modalities for training a Faster R-CNN architecture with a ResNet-50 backbone. On the same dataset, IPHD, the authors in [21] created a three-channel image by concatenating two duplicated thermal images and one

depth image similar to a three-channel RGB image to make use of the ImageNet pre-trained weights for initialisation. They achieved similar results on the AP_{50} metric compared to [12]. In [9] the authors fused the data into a 3-channel image, combining the grey channel of each modality per channel and applying different weights to the thermal and optical channels by applying the addWeighted method of OpenCV[56]. This process involved a manual search to determine the optimal weighting scheme for the given data. To perform these unifications, the intensity values of the pixels from both images are multiplied by the desired weights and then added to compute the pixel intensity values of the resulting image. In this method, the thermal image pixel values are added to the optical image pixel values using weights ranging from 0.1 to 0.9 with a step size of 0.1.
The way a colour scheme is applied to translate the thermal to visual data plays a significant role, which is discussed in section 2.14.2 and 2.12.5. Further, the method is sensitive to spatial-temporal data misalignment among sensors, which can be caused by calibration errors, different sampling rates, or sensor defects. This sensitivity further highlights the limitations of early fusion in certain scenarios[50].

Feature level fusion (DL) Middle fusion can be seen as a compromise between early and late fusion. By combining feature representations from different sensing modalities at intermediate layers, this approach allows the network to learn cross-modal information with varying feature representations and depths. Authors in [50] argued that middle fusion is quite flexible, but that finding the "optimal" way to fuse intermediate layers for a specific network architecture can be challenging. This difficulty arises from the intricate interactions between features and the potentially vast array of possible fusion configurations. Nevertheless, merging feature representations from different sensing modalities at intermediate layers enables the network to learn cross-modalities with varying feature representations and depths. This fusion can occur at a specific layer only once or can be hierarchically fused, as depicted in Fig. 2.12, such as through deep fusion or 'short-cut fusion'. Based on [50], this figure further illustrates the intricate nature of middle fusion and the variety of approaches that can be taken to combine information from different sensing modalities.
Shortcut Fusion, as discussed in detail in the paper [92], is a technique employed in deep neural networks that involves creating additional pathways within the network. This allows early layers to directly contribute to later layers, aiming to combine the advantages of both early and late fusion. By utilising low-level feature fusion and high-level decision fusion, shortcut fusion has the potential to enhance accuracy. It preserves detailed information from earlier stages and incorporates it into the final decision stages. However, it's worth noting that this method may increase the complexity of the network, potentially requiring additional computational resources.
Deep Fusion, on the other hand, operates hierarchically at multiple levels within the network. This is beneficial in capturing intricate interactions between different modalities at an intermediate stage. The authors of [93] emphasise the importance of deep feature alignment for multi-modal object detection and how Deep Fusion improves detection accuracy and robustness against input corruptions and out-of-distribution data. Deep fusion allows for a more comprehensive understanding of the data, as it integrates information from various stages of processing. Furthermore, deep middle fusion is often favoured over late fusion due to its superior feature integration capabilities. By combining features at a deeper level, it can lead to a more robust and reliable model, thereby enhancing the overall performance. The complexity of middle fusion is further illustrated by the variety of techniques that can be implemented, as identified by the authors in [94]. These techniques include:

- **Additive Fusion:** Individual networks or branches process each sensing modality up to a designated intermediate layer. The feature maps from these intermediate layers are either added element-wise or concatenated. The resulting feature maps are further processed by the network to produce the final output.
- **Multiplicative Fusion:** Separate networks or branches handle each sensing modality up to a specific intermediate layer. The feature maps from these intermediate layers are multiplied element-wise. The combined feature maps undergo further processing within the network to generate the final output.
- **Skip Connections:** Separate networks or branches process each sensing modality. Feature representations from intermediate layers are combined via skip connections. The combined feature maps are further processed by the network to produce the final output.

Drawing on the insights from the authors of [50] and [94], it can be stated that Shortcut Fusion and Deep Fusion serve as overarching frameworks for integrating features from diverse modalities. Within these frameworks, specific techniques such as Additive Fusion, Multiplicative Fusion, and Skip Connections can be employed. Each of these

methods brings its own unique strengths and potential challenges to the table. Additive Fusion, Multiplicative Fusion, and Skip Connections are distinct techniques that can be utilised to realise these fusion strategies. The choice between them depends on the specific requirements of the task at hand.
In [34], CNNs were used for feature extraction on the visual input of two modalities. In addition, hand-crafted feature extraction was applied to the thermal data for fusing these features in a Support Vector Machine (SVM) model. In their study, the authors asserted that face temperature variation contains significant differences between different actions, and thus can enhance the accuracy of activity recognition. They utilised synchronised thermal images to extract the face temperature variation of participants while they performed the actions. To achieve this, they manually selected the face region in the first frame and tracked it across frames using a KCF tracker[95]. Outliers in face temperature were removed when the movement was sudden or when the person was partially out of the frame. They divided the temperature values into 25 intervals and computed the average temperature for each interval. Then, they calculated the difference between every two consecutive intervals $(t_i - t_{i-1})$, resulting in 24 features that were added to the SVM model.
The authors in [24] extracted regions of interest (ROI) on the face using the Dlib[71] library to obtain the mean and variance of the temperatures in the ROIs as thermal features. The gait data of lower limbs were combined with these features for emotion detection, as lower limbs have more repeatable movements than the upper body. Joint angles and angular velocities were chosen as the features to characterize gait, including eight gait features based on the angle and velocity of the knees and hip. Convolutional Neural Networks (CNN), Hidden Markov Models (HMM), Support Vector Machines (SVM), and Random Forest (RF) models were employed to train and test the gait and thermal data. CNN and HMM models were trained with time series, whereas SVM and RF models were trained with static features such as the Power Spectral Density (PSD) of time series and the average temperatures of thermal image time series.

Decision level fusion (DL) Late fusion on the other hand offers high flexibility and modularity. When a new sensing modality is introduced, only the network associated with that modality needs to be trained, leaving the other networks unaffected. However, late fusion comes with drawbacks, such as increased computation costs and memory requirements. Additionally, it discards rich intermediate features that could be highly beneficial if fused, potentially limiting the performance of the overall system[50]. Late fusion in DL is commonly realised by the application of different versions of the non-maximum suppression algorithm (NMS) which works by first selecting the bounding box with the highest object detection score. Then, it compares the remaining bounding boxes and removes the ones that have a high degree of overlap as applied in [6]. [12] investigated the NMS method further and compared the Dual-NMS with the simple method and concluded that the Dual-NMS had a better performance. The Dual-NMS involves sorting two lists of detection boxes based on their confidence scores and collecting pairs from them. Similar to the basic NMS method, the boxes with the highest scores from each list are selected one by one and compared with the boxes from the other list. If a sufficient intersection over union (IOU) is found, the detection box is paired with the candidate with the highest score from the other list. The paired boxes are then merged into a single result, and the final detection box coordinates are updated through weighted averaging of the coordinates of the components[12].
A depiction of various fusion types is shown in Figure 2.12.

2.12.3 Fusion

The authors in [50] noted that they did not find definitive evidence that one fusion method would be superior to others based on their review of various methods using different stages. However, [12] and [21] conducted a comparison of the performance between the early middle and late fusion techniques and concluded that early fusion yielded better results in their use cases. These studies focused on the fusion of depth and thermal data for human detection, and it was discovered in [12] that using only depth data did not produce satisfactory detection outcomes. While the late fusion approach was slightly superior to using only depth data, it was inferior to using only the thermal data in comparison. However, [21] further argued that early fusion outperforms the other fusion methods in both the final detection results and computational complexity. Unlike late fusion, which only merges the detected boxes, early fusion enables cooperation between the depth and thermal information during feature extraction, allowing the model to extract and combine useful information from both modalities. It should be noted that while intermediate fusion also merges feature maps, the merging in early fusion is accomplished by a deep backbone network, leading to more effective cooperation between the depth and thermal information. In their study, they also demonstrate that the use of a Receptive Enhancement Module (REM) improves AP by 0.4, 0.9, and 2.3 at IOU thresholds of 0.25,

0.5, and 0.75, respectively. These findings suggest that the REM module enhances the accuracy of bounding box localisation but [12] achieved slightly better scores without it, which could be due to the slightly different fusion and training approach. In [21], the authors used a ResNet-50 backbone that was initialised with pre-trained parameters from ImageNet. However, the REM module and box prediction module were trained from scratch. [12] used the same backbone but models that were pre-trained using the Common Objects in Context dataset (COCO).
Besides constructing custom networks, some studies, such as [32], have utilised deep learning-based algorithms like OpenPose[96] to detect the pose of human occupants in a vehicle. Based on the derived bounding boxes this study applied late fusion. However, because OpenPose can only be applied to visual and thermal data, the authors did not utilise depth data. Other studies like [34] use CNNs for the feature extraction on the visual input of two modalities and apply a hand-crafted feature extraction for the thermal data to fuse these in a Support Vector Machine(SVM) model. However, they dropped the thermal feature in their experiments due to too much noise in the data.

2.12.4 Semantic Segmentation

Image semantic segmentation is a crucial task in computer vision, serving as an ideal perception solution for transforming image inputs into semantically meaningful regions and enabling pixel-wise dense scene understanding. Networks that rely solely on RGB information may face limitations in segmentation performance in complex environments or under challenging conditions. To enhance input information and improve performance, researchers have extensively explored multimodal sensor data fusion, which integrates additional data sources to provide a more comprehensive understanding of the scene. Various approaches have been proposed, such as FuseNet [94], which incorporates depth information, and HeatNet [97], which leverages thermal data for improved performance at night. Polarisation information has also been integrated into models, as seen in EAFNet [98]. Event data has been utilised in dense-to-sparse fusion to capture dynamic context information and improve segmentation performance, as in IS-SAFE [99]. Furthermore, there are specialised methods for RGB-D [100, 101, 102], RGB-T [103, 104, 105, 106] and RGB-P [15] semantic segmentation[15].
The authors in [107] argue that recently, vision transformers[89] have gained attention as they handle inputs as sequences and can acquire long-range correlations, providing a unified framework for diverse multi-modal tasks. But that multi-modal data often contain noisy measurements in different sensing modalities, such as low-quality distance estimation regions caused by limited effective depth ranges [108] and that compared to existing multi-modal fusion modules based on Convolutional Neural Networks (CNNs), it is not yet clear whether vision transformers can lead to significant improvements in RGB-X, where X stands for a different modality than RGB, semantic segmentation. Importantly, while some previous works like [108] and [109] use a simple global multi-modal interaction strategy, it may not generalise well across different sensing data combinations[110]. This is why the authors in [107] hypothesise that for RGB-X semantic segmentation with various supplements and uncertainties, comprehensive cross-modal interactions should be provided to fully exploit the potential of cross-modal complementary features why they propose CMX, a method designed to enhance semantic segmentation by incorporating diverse and complementary information from multiple modalities. CMX is a transformer-based cross-modal fusion framework that uses two streams to extract features from RGB images and the X-modality and includes a Cross-Modal Feature Rectification Module (CM-FRM) in each feature extraction stage to calibrate the feature of the current modality by combining the feature from the other modality. A Feature Fusion Module (FFM) is then used to mix the rectified feature pairs for the final semantic prediction. FFM includes a cross-attention mechanism, enabling the exchange of long-range contexts, and enhancing bi-modal features globally. Using a SegFormer-B2 backbone to visualise the segmentation results demonstrated that CMX improves the semantic segmentation of RGB-D data and identifies objects correctly, which are misclassified by the RGB-only model. For RGB-T segmentation, CMX provides clearer boundary distinctions between persons and unlabeled backgrounds in low illumination conditions. For RGB-P, CMX accurately segments specular glass areas, cars with polarisation cues, and pedestrians. For RGB-Event, CMX enhances the segmentation of moving objects. For RGB-LiDAR, CMX correctly segments the scene as compared to the RGB-only method. The results show that CMX is a suitable approach for multi-modal sensing combinations, providing robust semantic scene understanding[107]. The proposed CMX framework achieves state-of-the-art performances in different benchmarks but is limited to two simultaneous modalities at the time of writing.
Similar to [107], the authors in [15] put the focus on developing a generalisable multimodal perception system for various image modalities with an attention-based fusion architecture for outdoor scene understanding called NLFNet. This network is designed to effectively address the challenges of object segmentation in various complex scenarios.

The NLF (Non-Local Fusion) module, a key component of the network, is capable of adaptively extracting and fusing complementary information from different modal input images. It also leverages dependence information along with long-range contextual and positional priors to enhance the accuracy of semantic segmentation and applies a weighting mechanism based on a sigmoid activation function for fusing the modalities. By addressing these challenges, NLFNet aims to improve the performance of outdoor scene understanding across a range of conditions and input modalities. The network architecture is inspired by efficient networks such as SwiftNet [111] and RFNet [112]. NLFNet uses an encoder-decoder structure and adopts a ResNet-18 [113] backbone for each of its two independent branches. The encoder extracts latent features from RGB and other modal images, which are then merged using fusion operations. The Spatial Pyramid Pool (SPP) module [114], [112] is employed to expand effective receptive fields and generate feature maps with more global contextual information.

NLFNet incorporates efficient upsampling modules from SwiftNet [111] and merges RGB branch information through skip connections, improving segmentation accuracy. The Non-Local Fusion (NLF) module, inspired by Non-Local block [115] and NANet [101], integrates complementary information from RGB and other branches for the multi-level fusion of feature maps. The NLF module consists of two sub-modules: the Spatial Dependency Module (SDM) and the Channel Dependency Module (CDM).

The SDM establishes long-range contextual dependency between RGB and other modal branches in space, using global average pooling and convolutions to expand receptive fields. The CDM concatenates outputs from the SDM module along the channel dimension, obtaining a merged feature map, and performs global average pooling to obtain a squeezed feature map. It then adaptively transforms these embeddings into dependency weights via a sigmoid activation layer. This process establishes non-local contextual dependencies between different modalities and extracts nonlinear interactions between cross-modal channels.

The authors of the study demonstrate the effectiveness and generalisation ability of NLFNet across various multimodal sensor combinations. By conducting experiments with different sensor data, such as RGB-Depth, RGB-Polarisation, and RGB-Thermal images, they showcase the ability of NLFNet to handle diverse modalities and effectively fuse the complementary information. The results indicate that NLFNet is capable of providing accurate semantic segmentation in various challenging scenarios, proving its potential as a robust solution for outdoor scene understanding. But like CMX [107], the solution is bi-modal only.

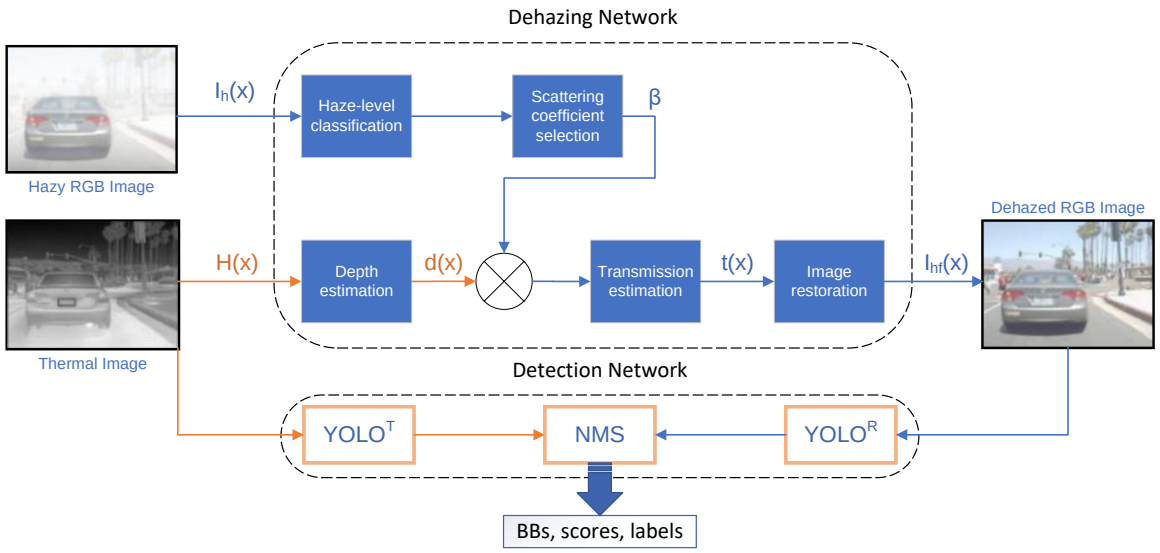


Figure 2.13: The overall architecture of the deep multimodal detection strategy. Object detection using YOLO from clear RGB images with rich colour information and thermal images with clear object bounding lines. The model detects the object with the highest probability through late fusion. [6]

In [4] applied the Residual Encoder-Decoder Network (RedNet) [116], which is a high-performing semantic segmentation network proposed in [102] that improves segmentation results by incorporating depth information into RGB signals. RedNet utilises an encoder-decoder network structure[117] with residual blocks as building modules, as well as a pyramid supervision training scheme to optimise the network. The encoder structure includes two convolutional branches, one for RGB and one for depth, that have the same configuration except for the feature channel number

of the convolution kernel, and feature fusion is achieved through element-wise summation. During training, the dataset was augmented and stochastic gradient descent (SGD) was used to optimise the network parameters with an initial learning rate of 0.002. The model is capable of segmenting target mask images from the background through inference. The authors augmented the training dataset by applying random scale and crop, followed by random hue, brightness, and saturation adjustment, which increased the dataset from 600 to 60,000 groups. Their model converged after approximately 100 epochs of training.

2.12.5 Object Detection

There are two main types of object detection algorithms that utilise convolutional neural networks (CNNs): two-stage detectors and single-stage detectors. The R-CNN family is a popular example of two-stage detectors, which typically use region-based methods. One such version is Faster R-CNN, used in [12], which introduced the region proposal network (RPN). The RPN can predict both the bounding box and the score at each position simultaneously, leading to a significant decrease in prediction time. An example of a popular single-stage network is YOLO[72]. Most studies, [9][6][37], utilise these networks with slight adjustments and perform early data level fusion for RGB-DT or just RGB-D or RGB-T object detection.

Only a single study, [5], focused on RGB-DT data fusion for Salient Object Detection (SOD) and implemented feature-level fusion with a CNN. In this work, the VGG16 classification network is used as a backbone for feature extraction. The tri-modal images are encoded separately using a three-stream encoding network, which extracts five-level features with varying resolutions. The authors proposed a hierarchical weighted suppress interference(HWSI) method to achieve an effective fusion of cross-modal information while also suppressing interference. The approach taken can be classified as a middle fusion with skip connections. This method involves assigning weights to each modality based on their importance for the given task and then using these weights to selectively suppress the interference introduced by each modality. By hierarchical weighting and selectively suppressing interference, the HWSI method can effectively fuse the cross-modal information but comes with a high computational cost which makes it less suitable for real-time applications. The HWSI method is composed of three distinct modules: the dual-modal attention fusion module (DMAFM), the triple-modal interactive weighting module (TMIWM), and the global attention-weighted fusion module (GAWFM). Each module is specifically designed to employ cross-modal information weighting to emphasize the salient regions and suppress interference effectively. The feature extraction is achieved by applying atrous convolutions with different dilation rates which can improve the performance of the network in tasks such as image segmentation and object detection. This approach, however, relies on the visual representation of the thermal and depth modality and some of the limitations that affect them are discussed in Section 2.14.2. The dataset created by the authors of [5] limits the thermal and depth representation to 256 values and a dynamic colour AGC algorithm is applied to the thermal data, as shown in Figure 2.19 and no gain control is applied to the depth data as shown in Figure 2.21. This can reduce the performance of object detection.

In contrast to the previous study, the authors of [6] utilised late fusion with two separate YOLO[72] models, following the thermal data dehazing process discussed in the Process Support2.9 section. The overall system architecture is illustrated in Figure 2.13. After completing the dehazing process, the resulting $I_{hf}(x)$ and $H(x)$ are fed into two YOLO models, denoted as $YOLO^R$ and $YOLO^T$, which use the image of their respective modality. Non-maximum suppression (NMS) is then employed to achieve late fusion. Their proposed model also allows for an RGB image with improved quality, as some of the haze removal can be performed using the haze level estimates. Moreover, by using late fusion and thermal images, the proposed model can process the rich colour and clear boundary information from both the RGB and thermal images simultaneously.

[9] investigated using a pre-trained YOLOv4 network on the COCO dataset, as well as training YOLOv4 on their own dataset. They limited the scope to human detection only but investigated different ways of fusing the images in an early fusion as discussed here 2.12.2. As their proposed data level fusion comes at a lower computational cost (average of 20ms. per frame) and the single-stage object detectors are fast, this system could be applied to real-time problems such as surveillance similar to study [12]. However, neither study published data related to the inference speed, and while [12] used a public dataset that was temporally aligned, [9] did not discuss temporal sensor alignment and used a thermal camera that was limited to 8 FPS.

In [37], the authors employed all three modalities, RGB, depth, and thermal data, to detect hands using a YOLO-based object detection algorithm [72] for real-time performance. Their analysis of the results suggested that using 2D bounding box detection with all three modalities led to higher accuracy compared to state-of-the-art model-based

RGB-D hand pose detection algorithms. Additionally, the authors found that RGB and thermal data were the most crucial modalities for this task. To train the YOLO detector, pre-trained features on ImageNet [118] were used, and the annotated bounding boxes in the dataset were employed for training. Since pre-training is only available for RGB images, knowledge distillation [119] was used to transfer pre-trained features to the thermal and depth modalities. Overall, deep learning-based approaches for multi-modal sensor fusion have shown promising results in various applications, and their continued development is expected to significantly advance the capabilities of multi-modal sensing systems in the future.

2.12.6 Presentation Attack Detection (PAD)

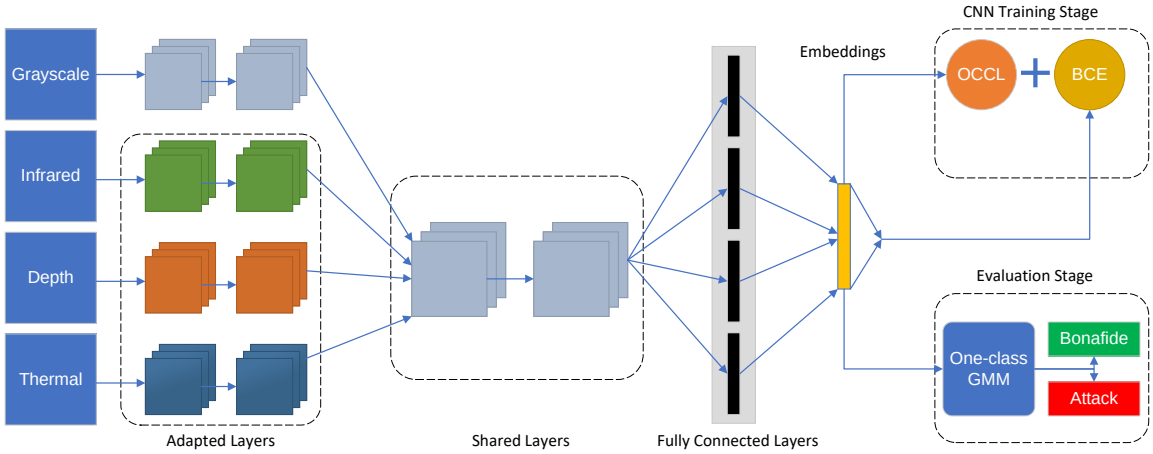


Figure 2.14: The CNN architecture is trained with two losses, the proposed One Class Contrastive Loss (OCCL) and Binary Cross Entropy (BCE), and then used as a fixed feature extractor with frozen weights. The one-class Gaussian Mixture Model (GMM) is trained using the embeddings obtained from the bona fide class alone. [26]

Biometrics provides a secure and convenient method for access control. Among various biometric modalities, face biometrics is one of the most preferred due to its non-intrusive nature. Despite the high performance of systems in identifying individuals in many challenging datasets, they are still vulnerable to presentation attacks (PA). It was identified that PAD in visual spectra alone is insufficient for security-critical applications. This area has seen a lot of attention in recent years and PAD systems are another area where fused RGB-DT data was applied. A multi-channel PAD framework called the Multi-Channel Convolutional Neural Network (MCCNN) was proposed in [30]. The MCCNN architecture is an extended version of the LightCNN model [120] adapted specifically for multi-channel PAD tasks and was then also applied in [26]. The main idea behind the MCCNN architecture is to leverage the joint

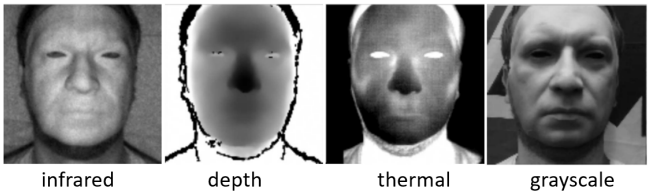


Figure 2.15: Preprocessed images resulting from a rigid mask attack [26].

representation from multiple channels for PAD tasks, using a pre-trained face recognition network. In this approach, a pre-trained LightCNN face recognition model is extended to accept multiple channels, and the embeddings from all channels are concatenated. Two fully connected layers are added on top of this joint representation layer for the PAD task. The first fully connected layer has ten nodes, and the second layer has only one output node. The higher-level features in the LightCNN part are shared among all modalities. The advantage of this architecture is that only lower layer features, known as Domain Specific Units (DSUs) [26], and higher-level fully connected layers are adapted in the training phase. This approach has two main advantages: first, a smaller number of parameters since the high-level features are shared across modalities, and second, adapting only DSUs and the final fully connected layers reduces

possible over-fitting since PAD databases are typically small in size. An optimal set of layers to be adapted was obtained empirically and was used in the baseline MCCNN and the proposed approach. Figure 2.15 shows a set of preprocessed images and Figure 2.14 the Schematic diagram of the proposed system found in [26].
In [26], the authors proposed a framework that utilises a one-class classifier along with a novel loss function, which encourages the CNN to learn a compact yet discriminative representation for face images.
As part of study [30], a publicly available dataset called The Wide Multi-Channel Presentation Attack (WMCA) database, was released. [27] made a similar dataset with higher quality and more modalities available which was called HQ-WCMA. Three studies [30, 26, 27] were co-written by some of the same authors who further developed their

Table 2.3: Publicly available RGB-DT datasets.

Study	Name	Type	Modalities	Link
[5]	VDT-2048	General	RGB D T	VDT-Dataset
[39]	VAP Trimodal People Seg.	People	RGB D T	VAP TPS-Dataset
[121]	KAIST	Driving	RGB D* T	KAIST-Dataset *
[30, 26]	WMCA	Faces/Masks	RGB D T	WMCA-Dataset
[27]	HQ-WMCA[122]	Faces/Masks	RGB D T	HQ-WMCA-Dataset
[7]	TriModal Face Detection Dataset	Faces	RGB D T	TMFD-Dataset

* Includes LiDAR and stereo RGB images.

Table 2.4: Publicly available Bi-Modal datasets used in studies.

Study	Name	Type	Modalities	Link
[12, 21]	IPHD[123]	People	D T	IPHD-Dataset
[107]	NYU Depth V2[124]	Indoor	RGB D	NYU Depth V2-Dataset
[107]	SUN-RGBD[125]	Indoor	RGB D	SUN-RGBD-Dataset
[107]	Stanford2D3D[126]	Indoor	RGB D	Stanford2D3D-Dataset
[107]	ScanNetV2[127]	Indoor	RGB D	ScanNetV2-Dataset
[107]	Cityscapes[128]	Driving	RGB D	Cityscapes-Dataset
[107]	MFNet[92]	Driving	RGB T	MFNet-Dataset
[107]	EventScape[129]	Driving	RGB E	EventScape-Dataset
[107]	KITTI-360[130]	Driving	RGB L	KITTI-360-Dataset

ideas in [27]. The data was collected using a custom-made sensor suite that enabled the recording of both genuine faces and presentation attacks across five different image modalities, including RGB, NIR, SWIR, thermal, and depth. Moreover, four banks of six LED modules were employed for illumination, providing coverage in 10 different wavelengths ranging from 735nm to 1650nm, encompassing the NIR and SWIR spectra. Sequential switching of these infrared emitters, synchronised with camera exposure periods, yielded multi-spectral reflectivity data across the sample. These wavelengths were chosen to provide the best possible multi-spectral coverage given market availability, resulting in 14 different modalities in each recording, including four NIR and seven SWIR wavelengths. The cameras were co-registered using a calibration procedure, enabling the captured data to be aligned in each modality. Experimental results showed that the investigated CNN models with SWIR outperformed baselines when a wide variety of attacks was considered, with almost perfect detection of all impersonation attacks while maintaining a low BPCER. However, the generalisation ability of the models using SWIR data was assessed on a cross-database experiment, revealing a noticeable difference on bona fide attempts, suggesting future research directions.
The proposed database 2.3 and code for studies [26][1] and [27][2] for reproducing the experiments are freely available for research purposes.

[1] https://gitlab.idiap.ch/bob/bob.paper.oneclass_mccnn_2019
[2] https://gitlab.idiap.ch/bob/bob.paper.pad_mccnns_swirdiff

2.13 Datasets

The majority of studies examined in this paper faced a scarcity of publicly available datasets, leading them to develop their own. Although some studies claimed to make their datasets public, like the TriModal Face Detection dataset(TMFD) [7], it could not be found during the writing of this survey. However, subsequent to the preprint release of our paper, we were contacted by the authors of the TMFD, who have now made their dataset publicly available. This dataset is a comprehensive resource that encompasses a wide range of variations, including different numbers of people in the scene, various backgrounds and distances. The dataset is categorised into three separate groups based on the complexity and difficulty level of face detection. Other studies, including [5] and [39], created and made their datasets available, and a list of all available tri-modal datasets can be found in Table 2.3. Public bi-modal datasets used by some reviewed papers are listed in Table 2.4. As this paper is centred on RGB-DT tri-modal fusion, only datasets used by the studies included in this review are presented.

2.14 Limitations

2.14.1 Sensors

One of the limitations of using thermal cameras in conjunction with RGB-D cameras is the potential mismatch in their respective field of view (FOV) and focal length, which can restrict the effective distance between objects or subjects being monitored. This discrepancy can result in inconsistencies in the size, position, and orientation of objects in the captured images, which can impact the accuracy of object detection and tracking algorithms. Another thermal sensor limitation is the need for Non-Uniformity Correction (NUC). NUC compensates for inconsistencies in the sensor's response to temperature changes, which can lead to inaccuracies in temperature measurements. This correction is required periodically and involves a mechanical shutter operation that blocks the imaging sensor with a material of uniform temperature for a short time up to a second.
The authors in [121] identified limitations with the temporal alignment in capturing images simultaneously using multiple devices. Despite the use of a signal generator to match the shutter times between devices, there can still be drift due to differences in exposure times. This can lead to an asynchronous phenomenon, especially during excessive movement of a vehicle or object.
It is important to note that temporal alignment is a critical factor in multi-camera systems, as it ensures that the images captured by different cameras are synchronised and can be properly used in applications such as 3D reconstruction or object detection.

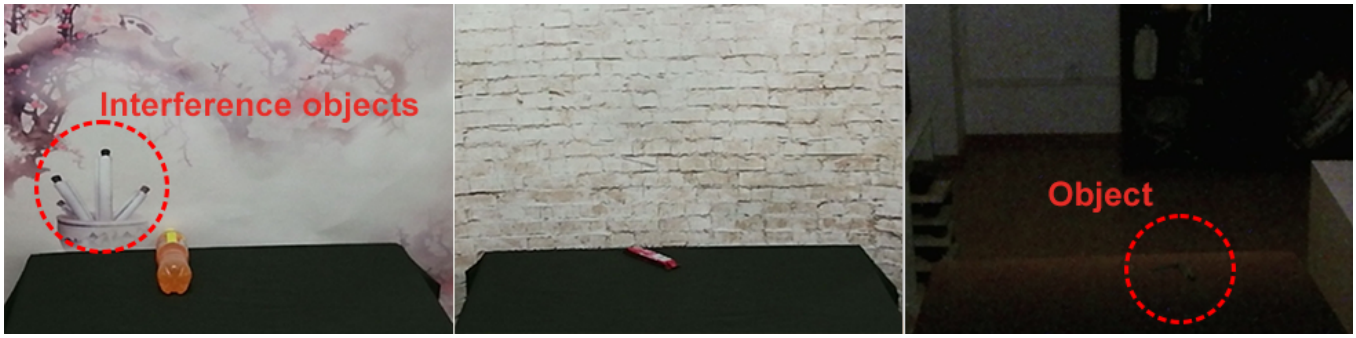
2.14.2 Modalities

In [5], the authors identified that visual perception systems that rely solely on RGB cameras face challenges such as:

1. The objects to be recognised in indoor environments are often small, numerous, dense, and vulnerable to background interference.

2. In low-light conditions, the ability to detect objects is greatly reduced, as illustrated in Figure 2.16.

To overcome the above problems, thermal and depth modalities can be introduced but despite that those sensors improve the detection of salient objects, they also introduce interference challenges and have their own individual challenges as can be seen in Figure 2.17 and Figure 2.18.
Figure 2.17(a) illustrates that the background of the depth image without any salient objects is very cluttered, which can distract the detection of salient object detection algorithms. Also, the depth information of a salient object can be incomplete when there is no distance difference between it and the surrounding objects, or when the difference is minimal. Furthermore, depth sensing can still be challenging for detecting some small objects.
Thermal sensors also present several challenges that need to be addressed, including thermal crossover, thermal radiation dispersion, and heat reflection. Thermal crossover occurs when the temperature of a salient object is the same as that of a portion of the background, as illustrated in 2.18(a), greatly increasing the difficulty of object detection. Figure 2.18(b) demonstrates an example of thermal radiation dispersion, where a portion of a salient object appears



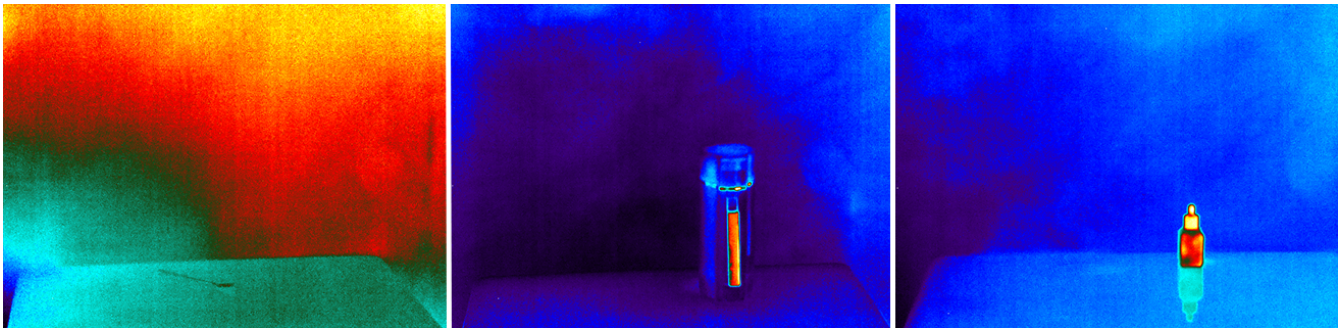
(a) (b) (c)

Figure 2.16: RGB modality challenges. (a) Similar appearance. (b) Small salient object. (c) Low illumination. [5]

(a) (b) (c)

Figure 2.17: Difficult challenges of depth images. (a) Background messy. (b) Depth information is incomplete. (c) Small salient objects. [5]

more salient than the rest of the object, causing interference to detection. Additionally, some objects exhibit heat reflection phenomena, as shown in 2.18(c), which is another important interference that needs to be addressed.



(a) (b) (c)

Figure 2.18: Taken from the VDT-2048 dataset, demonstrating the three identified thermal challenges: (a) Thermal crossover, (b) Thermal radiation dispersion, (c) Heat reflection. [5]

Accurate assignment of thermal values, as identified in [45], requires careful attention due to the nature of thermal-infrared sensors. As regular NUCs are required, real-time systems must be able to cope with thermal data interruption. If the correctly measured temperature is of importance, it should be considered that the thermal value can also be affected by the incident angle between the sensor and radiation emitted from the surface. Minimising this angle is considered best practice. The authors of [45] suggested three possible strategies to mitigate this:

- Perform Non-Uniformity Corrections (NUCs) more frequently, approximately every minute.
- Disregard frames obtained from the camera while a NUC is in progress.
- Assign temperatures only to rays with an incident angle of less than 30 degrees.

2.15 Synthesis

This section is intended to synthesise the key findings from our extensive review of the literature on the fusion of RGB-DT sensor modalities. It encapsulates the current state of the art, summarising the significant advancements and methodologies in this field across various applications. Before we transition into discussing the challenges, future work, and conclusion, this synthesis serves as a succinct recapitulation of the key points, including some insightful observations.

The traditional approach for the geometric calibration of thermal cameras, using a printed chessboard and a flood lamp, was inaccurate and difficult to execute. Geometric masks with high thermal contrast were introduced as an alternative calibration pattern, and multi-material calibration boards made of two materials with different emissivities have been developed for cross-calibrating thermal and visual modalities which have proven to be reliable and accurate. The registration of the modalities is applied based on the processing requirements, with offline approaches utilising computationally intensive feature-point matching algorithms. One widely used technique, especially for large-scale 3D reconstruction, is the Bundle Block Adjustment (BBA) with current advancements and improvements to the basic approach, such as using more advanced optimisation algorithms (e.g. Levenberg-Marquardt[131], Gauss-Newton[132]) [10]. In real-time 3D reconstruction, thermal data was added to the back-end of SLAM systems to enhance the robustness under unstable illumination environments and research in this area focuses on improving real-time performance by further reducing computational time and offering better model quality. In contrast, real-time processing for semantic segmentation and object detection requires performing geometric image rectification and alignment as a preprocessing step to ensure correct feature extraction. While studies for multi-modal semantic segmentation of RGB, depth and thermal data based on recent transformer networks were found [107, 15], those works only processed two modalities at a time and future research is aimed at processing more modalities simultaneously. However, the two studies demonstrate how to fuse bi-modal in real-time successfully and [15] used an adaptive weighting of modalities with a sigmoid activation layer to limit interference. The majority of the reviewed papers on multi-modal object detection utilised early fusion to generate a fused 8-bit three-channel image. Common object detectors such as YOLO were employed for the detection task, and some studies incorporated late fusion with the bounding boxes acquired from individual streams. A single study, [5], was identified that fused all three modalities (RGB, depth, and thermal) in a neural network using VGG16 as the feature extraction backbone. The study incorporated an interference suppression module to weigh the feature information across different modalities, mitigating interference from a single modality and compensating for potential information gaps in certain modalities. However, the computational requirements of this approach rendered it unsuitable for real-time processing. This study additionally created a publicly available, generic tri-modal dataset. Apart from this dataset, there are only three other RGB-DT datasets, which are specialised in presentation attack detection(PAD) and human detection applications. Another area that attracted a lot of attention is PAD where researchers focused on the problem of generalisation of the system. The challenge was addressed in [26] by building upon the Multi-Channel CNN (MCCNN) originally proposed in [30]. The authors developed a one-class classifier framework that employs learned features and a new loss function. This innovative loss function compels the CNN to acquire a concise and discriminative representation of face images, which enhances the overall performance even when used with the RGB modality alone. The authors showcased that their CNN method surpasses existing state-of-the-art feature-based techniques, while future research will focus on addressing the issue of potential attackers attempting to impersonate others.

2.16 Challenges and future work

2.16.1 Data Fusion

When it comes to fusing different modalities, one of the main challenges is sensor calibration and registration, especially when the sensors have different fields of view (FOV) that can cause parallax. To address this, techniques such as geometric calibration and image registration can be used to align the data from different sensors and reduce the effects of parallax. However, these techniques can be complex and time-consuming and still not produce perfect results, which can affect the accuracy of the fused data. Aside from employing software solutions to rectify parallax, another alternative involves using a beam-splitter, which enables two cameras to view the scene from the same point and utilise similar lenses to minimise parallax effects. Eliminating misalignment between modalities is vital in early fusion, which is why sensor calibration and registration continue to be significant research topics in this domain.

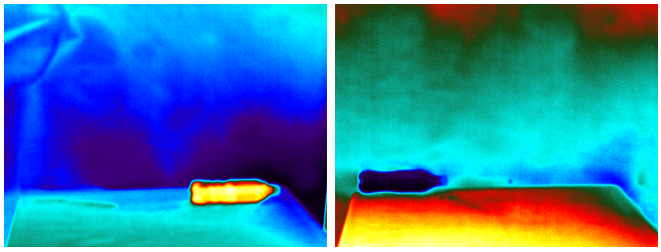


Figure 2.19: The image on the left shows a hot bottle on a table and the image on the right shows a cold bottle on the same table. Taken from the VDT-2048 dataset demonstrating the AGC colour shift of the same object(table) due to the application of a dynamic colour range based on the global minimum and maximum temperature in the frame.

2.16.2 Thermal Data

The mapping of colours for the display of thermal data is a crucial element for systems that utilise thermal data in a visual form. However, there appears to be a shortage of discussion on this topic in the reviewed literature. Most cameras apply automatic gain control (AGC) which is based on the lowest and highest temperature at any given time, causing the colours to shift. Besides that, grayscale or colour images limited to 256 values are being used but it is not specified over what range of temperatures it is used for e.g. when monitoring a range of -20 to 120°C, the resolution would be 0.55°C. However, thermal cameras have a sensitivity expressed as Noise Equivalent Delta Temperature (NEdT), which can range from 0.020°C up to 0.075°C. It is crucial to consider both aspects when analysing thermal images as limiting the data to 8-bit discards detail. For example, the VDT-2048 dataset uses 256-colour thermal images with a dynamic thermal-to-colour range association, as shown in Fig.2.19. The adaptive AGC algorithm may be appropriate for some applications; however, it could lead to difficulties if the intensity or colour information is essential for feature extraction or used for network training. Furthermore, the dataset exhibits some distortion and parallax between the visual and thermal modalities, as illustrated in Figure 2.20. Notably, the KAIST [121] driving dataset features raw 14-bit thermal data; conversely, the pedestrian dataset only includes 8-bit data. It is important to mention that while a 14-bit sensor can represent values up to 16,383, in environments with ambient temperatures around 20°C, the raw data captured falls within a narrow band of the full range. As a result, compression and contrast enhancement is crucial for encoding thermal images. However, it is essential to recognise that enhancement operations in thermal images can artificially distort the data, causing the loss of the physical correlation between the radiant flux from infrared radiation and pixel intensity [65]. Besides the data preprocessing, authors in [5] identified thermal crossover, thermal radiation dispersion, and heat reflection as challenges when processing thermal data. It is believed that a thorough preprocessing of thermal data and addressing the identified challenges in the field of multi-modal fusion constitutes a relevant future research direction.

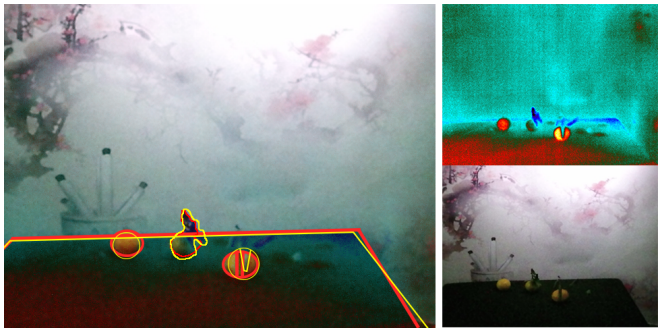


Figure 2.20: Taken from the VDT-2048 dataset, the image on the left shows an overlay of Visual(V) and Thermal(T) with some objects in T outlined in yellow and the same objects outlined in red in V. A parallax and distortion between the modalities can be observed. The images on the top right and on the bottom right are T and V, respectively.

2.16.3 Depth Data

Similar to thermal data, depth data is initially captured with a 16-bit resolution, but it is later converted to an 8-bit format when used as a depth map unlike point clouds, which are usually generated from the raw values. This conversion from 16-bit to 8-bit can lead to a loss of depth resolution and information due to the reduction in detail. Although this process can result in significant information loss, no studies in the reviewed literature have addressed this issue. It is essential to conduct thorough preprocessing of this modality when using it in the form of an 8-bit depth map. While the dynamically applied AGC algorithm in thermal images could cause issues, applying no processing at all will result in a loss of details. Figure 2.21 shows image 18 taken from the VDT-2048 dataset, in the original image on the left, it can be noted that visually almost nothing can be recognised as the observed depth is limited to a narrow band in the 16-bit data that was converted to 8-bit. On the right the same image with adjusted tonal balance by redistributing its brightness values. This is done by mapping the brightest and darkest pixel values in the image to white and black, respectively, and redistributing all the intermediate values evenly across the entire range.

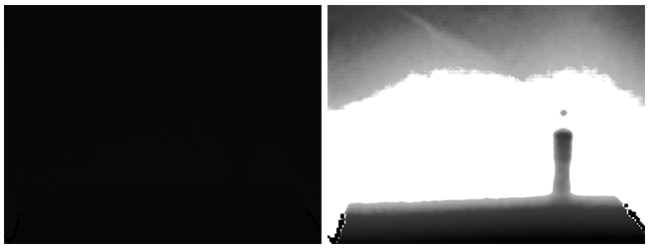


Figure 2.21: A comparative display of the original (left) and equalised (right) version of image 18 from the VTD-2048 dataset. The stark contrast between the two images accentuates the pivotal role of pre-processing in enhancing feature visibility, a critical step for the effective application of convolutional neural networks.

2.16.4 Datasets

The research potential in the field of tri-modal RGB-DT object detection is currently limited due to the lack of publicly available datasets. Apart from the VDT-2048 dataset mentioned earlier, there are no other tri-modal datasets suitable for general object detection, highlighting the need for more comprehensive datasets to advance research in this area.

2.16.5 Deep Learning

Integrating multiple modalities can improve object detection or segmentation accuracy, but it also increases computational requirements and makes real-time processing difficult. In the literature, studies show that early and late fusion of data has a relatively small impact on real-time performance; however, the potential for more sophisticated enhancements remains to be a challenge. More complex processing techniques applied in middle fusion, as demonstrated in [5] and [6], can result in frame rates dropping below 5 FPS. Even with the advances in Deep Neural Networks (DNNs) and hardware technology, achieving real-time object detection or segmentation with RGB-DT data still remains a challenge. While transformer-based architectures, as demonstrated in [107], and CNN-based architectures with non-local blocks, as demonstrated in [15], have shown promising results, they were limited to two concurrent modalities at the time of writing. Thus, further research is required to develop efficient and accurate fusion algorithms that can utilise three modalities and meet the requirements of real-time processing.

2.16.6 PAD

Currently, research in the field of PAD (Presentation Attack Detection) algorithms is centred on devising new methods capable of accurately detecting both known and unknown attacks. A major challenge for existing PAD algorithms is generalisation, as they often exhibit bias towards the training data. In [27], the authors recognised the SWIR (Short-Wave Infrared) spectrum, typically defined as light in the 0.9 – 1.7μm wavelength range but can also range from 0.7 – 2.5μm, as complementary and valuable. However, InGaAs (Indium Gallium Arsenide) sensor-based cameras for this spectrum are costly and reserved for specific applications. Meanwhile, in [26] the researchers tackled

the generalisation issue by building upon the Multi-Channel CNN (MCCNN) initially proposed in [30]. Their CNN approach with an innovative loss function outperformed all other methods. Future research will concentrate on addressing the challenge of potential attackers attempting to impersonate others.

2.17 Conclusion

This paper presents a comprehensive overview of the fusion between RGB-D and thermal modalities, exploring their applications, and the techniques employed. Over the past decade, there has been a surge of interest in fusing these modalities, demonstrating their considerable potential across diverse fields, including robotics, surveillance, medical imaging, and maintenance systems. Combining these modalities has proven to enhance the accuracy, robustness, and reliability of computer vision systems, contributing to the overall effectiveness of the technology. To systematically summarise the findings, a search strategy based on the PRISMA framework was used. The literature review has revealed several approaches for integrating RGB-D and thermal data, including feature-level fusion, decision-level fusion, and data-level fusion. Furthermore, the use of deep learning techniques has emerged as a popular approach for effectively combining RGB-D and thermal data, surpassing traditional feature-based approaches. Overall, the reviewed literature suggests that the fusion of RGB-D and thermal modalities holds great potential for enhancing the performance of computer vision systems in diverse applications and even creating new ones.
It was observed that researchers have primarily focused on the higher-level architecture of neural networks when conducting sensor fusion with deep learning while overlooking the importance of preprocessing steps. While Visual Transformers have shown promising results in sensor fusion, no existing tri-modal RGB-DT fusion has been developed thus far. Therefore, further research is necessary to develop advanced fusion techniques that can enhance the accuracy and reliability of the results while operating in real-time, thereby unlocking the full potential of this approach and making it applicable for various practical applications. In conclusion, this study aims to serve as a supplementary resource for researchers in the field of RGB-DT sensor fusion, providing a robust foundation and guidance for ongoing investigation and advancements.

Bibliography

[1] Waldemar Minkina and Sebastian Dudzik. *Infrared thermography: errors and uncertainties*. John Wiley & Sons, 2009.

[2] Karolj Skala, Tomislav Lipić, Ivan Sović, Luko Gjenero, and Ivan Grubišić. 4d thermal imaging system for medical applications. *Periodicum biologorum*, 113(4):407–416, 2011.

[3] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, AsbjÃ¸rn HrÃ³bjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1):89, 2021.

[4] Xiaomin Zhang, Yanning Zhang, Jinfeng Geng, Jinming Pan, Xinyao Huang, and Xiuqin Rao. Feather damage monitoring system using rgb-depth-thermal model for chickens. *Animals*, 13(1):126, 2023.

[5] Kechen Song, Jie Wang, Yanqi Bao, Liming Huang, and Yunhui Yan. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. *IEEE/ASME Transactions on Mechatronics*, 2022.

[6] Sungan Yoon and Jeongho Cho. Deep multimodal detection in reduced visibility using thermal depth estimation for autonomous driving. *Sensors*, 22(14):5084, 2022.

[7] Wiktor Mucha and Martin Kampel. Depth and thermal images in face detection-a detailed comparison between image modalities. In *2022 the 5th International Conference on Machine Vision and Applications (ICMVA)*, pages 16–21, 2022.

[8] Moritz Oppliger, Jonas Gutknecht, Roman Gubler, Matthias Ludwig, and Teddy Loeliger. Sensor fusion of 3d time-of-flight and thermal infrared camera for presence detection of living beings. In *2022 IEEE Sensors*, pages 1–4. IEEE, 2022.

[9] Ahmet Ozcan and Omer Cetin. A novel fusion method with thermal and rgb-d sensor data for human detection. *IEEE Access*, 10:66831–66843, 2022.

[10] Tanhao Zhang, Luyin Hu, Yuxiang Sun, Lu Li, and David Navarro-Alarcon. Computing thermal point clouds by fusing rgb-d and infrared images: From dense object reconstruction to environment mapping. In *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1707–1714. IEEE, 2022.

[11] Ruicheng Qiu, Yanlong Miao, Man Zhang, and Han Li. Detection of the 3d temperature characteristics of maize under water stress using thermal and rgb-d cameras. *Computers and Electronics in Agriculture*, 191:106551, 2021.

[12] Weronika Gutfeter and Andrzej Pacut. Fusion of depth and thermal imaging for people detection. *Journal of Telecommunications and Information Technology*, 2021.

[13] E Semenishchev, V Voronin, S Agaian, M Zhdanova, and A Zelensky. Algorithm for fusing data obtained by thermal, 3d, and the visible range cameras. In *Dimensional Optical Metrology and Inspection for Practical Applications X*, volume 11732, pages 105–111. SPIE, 2021.

[14] Yahya Zefri, Imane Sebari, Hicham Hajji, and Ghassane Aniba. In-depth investigation of applied digital photogrammetry to imagery-based rgb and thermal infrared aerial inspection of large-scale photovoltaic installations. *Remote Sensing Applications: Society and Environment*, 23:100576, 2021.

[15] Ran Yan, Kailun Yang, and Kaiwei Wang. Nlfnet: non-local fusion towards generalized multimodal semantic segmentation across rgb-depth, polarization, and thermal images. In *2021 IEEE international conference on robotics and biomimetics (ROBIO)*, pages 1129–1135. IEEE, 2021.

[16] Mario Ortega, Eugenio Ivorra, Alejandro Juan, Pablo Venegas, Jorge Martínez, and Mariano Alcañiz. Mantra: An effective system based on augmented reality and infrared thermography for industrial maintenance. *Applied Sciences*, 11(1):385, 2021.

[17] Yanpeng Cao, Yafei Dong, Fan Wang, Jiangxin Yang, Yanlong Cao, and Xin Li. Multi-sensor spatial augmented reality for visualizing the invisible thermal information of 3d objects. *Optics and Lasers in Engineering*, 145:106634, 2021.

[18] Haichao Zheng, Xue Zhong, Junru Yan, Lihua Zhao, and Xintian Wang. A thermal performance detection method for building envelope based on 3d model generated by uav thermal imagery. *Energies*, 13(24):6677, 2020.

[19] MaÅ,gorzata JarzÄ. . . bek-Rychard, Dong Lin, and Hans-Gerd Maas. Supervised detection of faÃ§ade openings in 3d point clouds with thermal attributes. *Remote Sensing*, 12(3):543, 2020.

[20] Imen Halima, Jean-Marc LafertÃ©, Geoffroy Cormier, Alain-JÃ©rÃ´me FougÃ¨res, and Jean-Louis Dillenseger. Depth and thermal information fusion for head tracking using particle filter in a fall detection context. *Integrated Computer-Aided Engineering*, 27(2):195–208, 2020.

[21] Zijian Zhao, Jie Zhang, and Shiguang Shan. Noise robust hard example mining for human detection with efficient depth-thermal fusion. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 809–813. IEEE, 2020.

[22] Farid Javadnejad, Daniel T Gillins, Christopher E Parrish, and Richard K Slocum. A photogrammetric approach to fusing natural colour and thermal infrared uas imagery in 3d point cloud generation. *International Journal of Remote Sensing*, 41(1):211–237, 2020.

[23] Yangyu Shi, Pierre Payeur, Monique Frize, and Erika Bariciak. Thermal and rgb-d imaging for necrotizing enterocolitis detection. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6. IEEE, 2020.

[24] Chuang Yu and Adriana Tapus. Multimodal emotion recognition with thermal and rgb-d cameras for human-robot interaction. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 532–534, 2020.

[25] Feng Zhao, Serhan Cosar, Nicola Bellotto, and Shigang Yue. Roi detection and tracking for physiological monitoring based on calibration between rgb-d and thermal cameras, 2020.

[26] Anjith George and SÃ©bastien Marcel. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 16:361–375, 2020.

[27] Guillaume Heusch, Anjith George, David Geissbühler, Zohreh Mostaani, and Sébastien Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

[28] Da Li, Carol Menassa, and Vineet R Kamat. Thermal and rgb-d sensor fusion for non-intrusive human thermal comfort assessment. In *CIB World Building Congress 2019, Hong Kong*, 2019.

[29] Xiaotian Chen, Guiyun Tian, Jianbo Wu, Chaoqing Tang, and Kongjing Li. Feature-based registration for 3d eddy current pulsed thermography. *ieee sensors journal*, 19(16):6998–7004, 2019.

[30] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, AndrÃ© Anjos, and SÃ©bastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2019.

[31] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019.

[32] Changho Song and Seung-Hun Kim. Robust vehicle occupant detection based on rgb-depth-thermal camera. *The Journal of Korea Robotics Society*, 13(1):31–37, 2018.

[33] David-Octavian Iacob and Adriana Tapus. First attempts in deception detection in hri by using thermal and rgb-d cameras. In *RO-MAN 2018*, 2018.

[34] Mihaela Sorostinean and Adriana Tapus. Activity recognition based on rgb-d and thermal sensors for socially assistive robots. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 1298–1304. IEEE, 2018.

[35] Ming-Der Yang, Tung-Ching Su, and Hung-Yu Lin. Fusion of infrared thermal image and visible image for 3d thermal model reconstruction using smartphone sensors. *Sensors*, 18(7):2003, 2018.

[36] Yanpeng Cao, Baobei Xu, Zhangyu Ye, Jiangxin Yang, Yanlong Cao, Christel-Loic Tisse, and Xin Li. Depth and thermal sensor fusion to enhance 3d thermographic reconstruction. *Optics express*, 26(7):8179–8193, 2018.

[37] Rachel Luo, Ozan Sener, and Silvio Savarese. Scene semantic reconstruction from egocentric rgb-d-thermal videos. In *2017 International Conference on 3D Vision (3DV)*, pages 593–602. IEEE, 2017.

[38] Ignacio Rocco Spremolla, Michel Antunes, Djamila Aouada, and BjÃ¶rn E Ottersten. Rgb-d and thermal sensor fusion-application in person tracking. In *VISIGRAPP (3: VISAPP)*, pages 612–619, 2016.

[39] Cristina Palmero, Albert ClapÃ©s, Chris Bahnsen, Andreas MÃ¸gelmose, Thomas B Moeslund, and Sergio Escalera. Multi-modal rgbâ€“depthâ€“thermal human body segmentation. *International Journal of Computer Vision*, 118:217–239, 2016.

[40] Wataru Nakagawa, Kazuki Matsumoto, Francois de Sorbier, Maki Sugimoto, Hideo Saito, Shuji Senda, Takashi Shibata, and Akihiko Iketani. Visualization of temperature change using rgb-d camera and thermal camera. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*, pages 386–400. Springer, 2014.

[41] Ramin Irani, Kamal Nasrollahi, Marc O Simon, Ciprian A Corneanu, Sergio Escalera, Chris Bahnsen, Dennis H Lundtoft, Thomas B Moeslund, Tanja L Pedersen, Maria-Louise Klitgaard, et al. Spatiotemporal analysis of rgb-dt facial images for multimodal pain level recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 88–95, 2015.

[42] Johannes Rangel, Julian GarzÃ³n, Jorge Sofrony, and Andreas Kroll. Gas leak inspection using thermal, visual and depth images and a depth-enhanced gas detection strategy. *Revista de Ingeniería*, 42(1):8–15, 2015.

[43] Stephen Vidas, Peyman Moghadam, and Sridha Sridharan. Real-time mobile 3d temperature mapping. *IEEE Sensors Journal*, 15(2):1145–1152, 2014.

[44] J Rangel, S Soldan, and A Kroll. 3d thermal imaging: Fusion of thermography and depth cameras, robotics and autimation. In *IEEE Intern. Conf. on Quantitative InfraRed Thermography*, pages 2311–2318, 2014.

[45] Stephen Vidas, Peyman Moghadam, and Michael Bosse. 3d thermal mapping of building interiors using an rgb-d and thermal camera. In *2013 IEEE international conference on robotics and automation*, pages 2311–2318. IEEE, 2013.

[46] Loreto Susperregi, Jose Maria MartÃnez-Otzeta, Ander Ansuategui, Aitor Ibarguren, and Basilio Sierra. Rgb-d, laser and thermal sensor fusion for people following in a mobile robot. *International Journal of Advanced Robotic Systems*, 10(6):271, 2013.

[47] Andreas Mogelmose, Chris Bahnsen, Thomas Moeslund, Albert ClapÃ©s, and Sergio Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 301–307, 2013.

[48] Dorit Borrmann, Andreas Nüchter, Marija Đakulović, Ivan Maurović, Ivan Petrović, Dinko Osmanković, and Jasmin Velagić. The project thermalmapper–thermal 3d mapping of indoor environments for saving energy. *IFAC Proceedings Volumes*, 45(22):31–38, 2012.

[49] Youngjib Ham and Mani Golparvar-Fard. Rapid 3d energy performance modeling of existing buildings using thermal and digital imagery. In *Construction Research Congress 2012: Construction Challenges in a Flat World*, pages 991–1000, 2019.

[50] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[51] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[52] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[53] Stephen Vidas, Ruan Lakemond, Simon Denman, Clinton Fookes, Sridha Sridharan, and Tim Wark. A mask-based approach for the geometric calibration of thermal-infrared cameras. *IEEE Transactions on Instrumentation and Measurement*, 61(6):1625–1635, 2012.

[54] Namil Kim, Yukyung Choi, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, and In So Kweon. Geometrical calibration of multispectral calibration. In *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 384–385. IEEE, 2015.

[55] Tanhao Zhang, Luyin Hu, Lu Li, and David Navarro-Alarcon. Towards a multispectral rgb-ir-uv-d vision systemâ€"seeing the invisible in 3d. In *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1723–1728. IEEE, 2021.

[56] G. Bradski and A. Kaehler. OpenCV library. `https://opencv.org/`, 2000. Accessed: March 22, 2023.

[57] Gil Ben-Artzi, Tavi Halperin, Michael Werman, and Shmuel Peleg. Epipolar geometry based on line similarity. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1864–1869. IEEE, 2016.

[58] R Istenic, D Heric, S Ribaric, and Damjan Zazula. Thermal and visual image registration in hough parameter space. In *2007 14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services*, pages 106–109. IEEE, 2007.

[59] The MathWorks Inc. MATLAB. `https://www.mathworks.com/products/matlab.html`, 2022. Accessed on March 22, 2023.

[60] Karl Kraus. *Photogrammetry: Geometry from Images and Laser Scans*. Walter de Gruyter, 2 edition, 2007.

[61] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[62] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. *Lecture notes in computer science*, 3951:404–417, 2006.

[63] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 International conference on computer vision*, pages 2548–2555. Ieee, 2011.

[64] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J Davison. Kaze features. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 214–227. Springer, 2012.

[65] Maria João Sousa, Alexandra Moutinho, and Miguel Almeida. Thermal infrared sensing for near real-time data-driven fire detection and monitoring systems. *Sensors*, 20(23):6803, 2020.

[66] Jian Nie, Jun Yan, Huilin Yin, Lei Ren, and Qian Meng. A multimodality fusion deep neural network and safety test strategy for intelligent vehicles. *IEEE transactions on intelligent vehicles*, 6(2):310–322, 2020.

[67] Mircea Paul Muresan, Ion Giosan, and Sergiu Nedevschi. Stabilization and validation of 3d object position using multimodal sensor fusion and semantic segmentation. *Sensors*, 20(4):1110, 2020.

[68] Marcin Kowalski and Krzysztof Mierzejewski. Detection of 3d face masks with thermal infrared imaging and deep learning techniques. *Photonics Letters of Poland*, 13(2):22–24, 2021.

[69] Dae-Eon Kim, BongKyu Jeon, and Dong-Soo Kwon. 3d convolutional neural networks based fall detection with thermal camera. *The Journal of Korea Robotics Society*, 13(1):45–54, 2018.

[70] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 354–361, 2013.

[71] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[72] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[73] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979.

[74] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011.

[75] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017.

[76] Thomas Whelan, Michael Kaess, Hordur Johannsson, Maurice Fallon, John J Leonard, and John McDonald. Real-time large-scale dense rgb-d slam with volumetric fusion. *The International Journal of Robotics Research*, 34(4-5):598–626, 2015.

[77] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslamâ€"learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2560–2568, 2018.

[78] Martin Landmann, Stefan Heist, Patrick Dietrich, Peter Lutzke, Ingo Gebhart, Joachim Templin, Peter Kühmstedt, Andreas Tünnermann, and Gunther Notni. High-speed 3d thermography. *Optics and Lasers in Engineering*, 121:448–455, 2019.

[79] Junhua Sun, Haining Ma, and Debing Zeng. Three-dimensional infrared imaging method based on binocular stereo vision. *Optical Engineering*, 54(10):103111–103111, 2015.

[80] The Open3D development team. Open3d: A modern library for 3d data processing, 2023. Version 0.13.0. Available at: `http://www.open3d.org/`.

[81] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003.

[82] Jose Maria Martínez-Otzeta, Aitor Ibarguren, Ander Ansuategi, and Loreto Susperregi. Laser based people following behaviour in an emergency environment. In *Intelligent Robotics and Applications: Second International Conference, ICIRA 2009, Singapore, December 16-18, 2009. Proceedings 2*, pages 33–42. Springer, 2009.

[83] Nicola Bellotto, Huosheng Hu, et al. A bank of unscented kalman filters for multimodal human perception with mobile service robots. *International Journal of Social Robotics*, 2(2):121–136, 2010.

[84] Kevin J Cannons and Richard P Wildes. The applicability of spatiotemporal oriented energy features to region tracking. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):784–796, 2013.

[85] Ramin Irani, Kamal Nasrollahi, and Thomas B Moeslund. Pain recognition using spatiotemporal oriented energy of facial muscles. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 80–87, 2015.

[86] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[87] Ozan Şener, Kemal Ugur, and A Aydın Alatan. Efficient mrf energy propagation for video segmentation via bilateral filters. *IEEE transactions on multimedia*, 16(5):1292–1302, 2014.

[88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[89] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[90] Yunfei Liu, Deng-Ping Fan, Ming-Ming Cheng, Tao Li, and Ali Borji. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1474–1483, 2019.

[91] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020.

[92] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[93] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022.

[94] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Computer Visionâ€“ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, pages 213–228. Springer, 2017.

[95] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014.

[96] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1302–1310, 2017.

[97] Johan Vertens, Jannik ZÃ¼rn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8461–8468. IEEE, 2020.

[98] Kaite Xiang, Kailun Yang, and Kaiwei Wang. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Optics Express*, 29(4):4802–4820, 2021.

[99] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. Issafe: Improving semantic segmentation in accidents by fusing event-based data. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1132–1139. IEEE, 2021.

[100] Hao Zhou, Lu Qi, Zhaoliang Wan, Hai Huang, and Xu Yang. Rgb-d co-attention network for semantic segmentation. In *Proceedings of the Asian conference on computer vision*, 2020.

[101] Guodong Zhang, Jing-Hao Xue, Pengwei Xie, Sifan Yang, and Guijin Wang. Non-local aggregation for rgb-d semantic segmentation. *IEEE Signal Processing Letters*, 28:658–662, 2021.

[102] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018.

[103] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020.

[104] Yeong-Hyeon Kim, Ukcheol Shin, Jinsun Park, and In So Kweon. Ms-uda: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. *IEEE Robotics and Automation Letters*, 6(4):6497–6504, 2021.

[105] Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2020.

[106] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2021.

[107] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023.

[108] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019.

[109] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pages 561–577. Springer, 2020.

[110] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2021.

[111] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12616, 2019.

[112] Lei Sun, Kailun Yang, Xinxin Hu, Weijian Hu, and Kaiwei Wang. Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images. *IEEE robotics and automation letters*, 5(4):5558–5565, 2020.

[113] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[114] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[115] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[116] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018.

[117] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[118] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[119] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[120] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[121] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[122] Zohreh Mostaani, Anjith George, Guillaume Heusch, David Geissenbuhler, and Sébastien Marcel. The high-quality wide multi-channel attack (hq-wmca) database. Idiap-RR Idiap-RR-22-2020, Idiap, 9 2020.

[123] Albert Clapés, Julio CS Jacques Junior, Carla Morral, and Sergio Escalera. Chalearn lap 2020 challenge on identity-preserved human detection: Dataset and results. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 801–808. IEEE, 2020.

[124] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012.

[125] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[126] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[127] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[128] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[129] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021.

[130] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[131] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical Analysis: Proceedings of the Biennial Conference Held at Dundee, June 28–July 1, 1977*, pages 105–116. Springer, 2006.

[132] F Dan Foresee and Martin T Hagan. Gauss-newton approximation to bayesian learning. In *Proceedings of international conference on neural networks (ICNN'97)*, volume 3, pages 1930–1935. IEEE, 1997.

Chapter 3

MM5: Dataset, Capture, Calibration and Preprocessing

The systematic review presented in Chapter 2 identified critical gaps that impede progress in multimodal fusion research: the scarcity of publicly available datasets extending beyond RGB-D-T configurations, the absence of benchmarks providing raw sensor data for depth and thermal modalities, limited support for evaluating both aligned and unaligned fusion paradigms, and insufficient attention to modality-specific preprocessing. This chapter directly addresses these gaps by introducing MM5, a five-modality dataset and processing pipeline that provides the empirical foundation for the fusion architectures developed in subsequent chapters. MM5 extends beyond the tri-modal scope surveyed in Chapter 2 by incorporating infrared intensity and ultraviolet imagery, preserving raw 16-bit measurements for depth, NIR intensity and thermal sensors, and introducing preprocessing algorithms (DTMRE, ADMRE) and annotation tools (MAR) that enable rigorous experimentation under both aligned and unaligned conditions.
The contents of this chapter are reproduced from the following article:

Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2025). MM5: Multimodal image capture and dataset generation for RGB, depth, thermal, UV, and NIR. Information Fusion*, 126, 103516. https://doi.org/10.1016/j.inffus.2025.103516*.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Martin Brenner
Name and title of main supervisor:	Dr Napoleon Reyes
In which chapter is the manuscript/published work?	3

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

The candidate was the main contributor of this work, and has done the literature review, experiments, and drafted the manuscript. The final draft was completed with the suggestions from the co-authors.

Please select one of the following three options:

◉ **The manuscript/published work is published or in press**

Please provide the full reference of the research output:

M. Brenner, N. H. Reyes, T. Susnjak and A. L. C. Barczak, "MM5: Multimodal image capture and dataset generation for RGB, depth, thermal, UV, and NIR," in Information Fusion, Volume 126, Part A, 2025, 103516, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2025.103516

○ **The manuscript is currently under review for publication**

Please provide the name of the journal:

○ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Digitally signed by Martin Brenner DN: cn=Martin Brenner, c=NZ, email=mb@fisag.co.nz Reason: I agree to specified portions of this document Location: Auckland Date: 2025.11.25 20:47:40 +13'00'	Main supervisor's signature:	Napoleon Reyes Digitally signed by Napoleon Reyes Date: 2025.12.01 17:09:17 +13'00'

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

3.1 Abstract

Existing multimodal datasets often lack sufficient modality diversity, raw data preservation, and flexible annotation strategies, seldom addressing modality-specific cues across multiple spectral channels. Current annotations typically concentrate on pre-aligned images, neglecting unaligned data and overlooking crucial cross-modal alignment challenges. These constraints significantly impede advanced multimodal fusion research, especially when exploring modality-specific features or adaptable fusion methodologies. To address these limitations, we introduce MM5, a comprehensive dataset integrating RGB, depth, thermal (T), ultraviolet (UV), and near-infrared (NIR) modalities. Our capturing system utilises off-the-shelf components, incorporating stereo RGB-D imaging to provide additional depth and intensity (I) information, enhancing spatial perception and facilitating robust cross-modal learning. MM5 preserves depth and thermal measurements in raw, 16-bit formats, enabling researchers to explore advanced preprocessing and enhancement techniques. Additionally, we propose a novel label re-projection algorithm that generates ground-truth annotations directly for distorted thermal and UV modalities, supporting complex fusion strategies beyond strictly aligned data. Dataset scenes encompass varied lighting conditions (e.g. shadows, dim lighting, overexposure) and diverse objects, including real fruits, plastic replicas, and partially rotten produce, creating challenging scenarios for robust multimodal analysis. We evaluate the effects of multi-bit representations, adaptive gain control (AGC), and depth preprocessing on a transformer-based segmentation network. Our preprocessing improved mean IoU from 70.66% to 76.33% for depth data and from 72.67% to 79.08% for thermal encoding, using our novel preprocessing techniques, validating MM5's efficacy in supporting comprehensive multimodal fusion research.

3.2 Introduction

The extraction and analysis of visual features using RGB cameras have been widely applied in computer vision across various industrial, commercial, and research domains. However, traditional RGB-based imaging is inherently limited by its confinement to the visible spectrum, making it highly dependent on external lighting conditions and susceptible to occlusions or variations in ambient illumination [1]. Expanding image capture to multiple modalities, such as depth (D), thermal (T), ultraviolet (UV), and near-infrared (NIR), can provide a more comprehensive understanding of a scene by leveraging different spectral characteristics. For example, near-infrared (NIR) imaging can penetrate certain materials or haze, and ultraviolet (UV) imaging can reveal surface details or substances not visible in RGB. Fusing such modalities with RGB yields a richer, more robust representation of the scene [2]. Studies show that multimodal combinations outperform single RGB; pairing thermal with colour imagery significantly improves pedestrian detection under difficult illumination conditions [3]. Motivated by such successes, research into multispectral perception has gained increasing attention recently due to its potential to enhance object recognition, segmentation, environmental monitoring, medical imaging, security, and robotics. Each sensor modality offers distinct advantages; for example, depth sensors provide 3D spatial information independent of texture, enabling enhanced perception of object shapes, positions, and distances; thermal cameras capture temperature differences between objects and their surroundings, ensuring reliable detection in low-light conditions and through occlusions; and UV and NIR imaging extend visibility beyond the human-perceivable spectrum, providing additional material and structural insights that can improve classification tasks. For instance, a recent comprehensive review details how the fusion of visible, NIR, and thermal imagery can dramatically improve the robustness of security systems for challenging tasks like biometric facial recognition [4], underscoring the broad scientific and industrial value of the sensor combination featured in our MM5 dataset. Recent advances in multimodal fusion have increasingly addressed the integration of three or more complementary sensing modalities to overcome the limitations of single-modality perception under challenging environmental conditions such as poor illumination, occlusion, and spectral camouflage. Early approaches primarily utilised convolutional neural networks (CNNs) designed for RGB, depth, and thermal infrared data, demonstrating improved robustness for salient object detection and segmentation [1, 5, 6, 7]. These models often employ separate backbones for each modality, incurring higher computational costs and complexity. More recent works have adopted transformer-based architectures and attention mechanisms for more efficient and flexible fusion of multiple modalities, including RGB, depth, intensity, thermal, and ultraviolet [8, 9, 10, 11]. Such models enable stage-wise or feature-level fusion with adaptive weighting, improving robustness against modality-specific noise and variability. Additionally, related research in remote sensing and cross-domain adaptation addresses challenges of generalising multimodal models across different geographic locations and environmental conditions, which are critical for practical

deployment in diverse real-world scenarios. However, despite the clear benefits of multimodal data fusion, research advancements in this domain are constrained by the limited availability of publicly accessible datasets that integrate multiple imaging modalities, such as RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR). As discussed in Section 3.3.1, many existing multimodal datasets predominantly target specific applications and frequently lack raw sensor data, resulting in restricted opportunities for exploring comprehensive multispectral fusion strategies. This limitation significantly hampers their broader applicability in multimodal fusion research. Moreover, some studies even rely on synthetic or artificially generated images [10], which fail to fully capture the complexities and challenges inherent in real-world multimodal data acquisition. Furthermore, accurate sensor calibration, precise alignment across modalities, and appropriate data preprocessing continue to pose considerable challenges, further complicating the effective utilisation of multimodal datasets in practical applications. To overcome these challenges, we introduce MM5 [12], a comprehensive multimodal dataset that systematically captures RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR) data, alongside stereo RGB-D imaging to provide complementary intensity and depth information. MM5's design specifically targets robust multimodal fusion research by offering raw 16-bit depth and thermal data, enabling researchers to apply their own denoising, enhancement, and preprocessing algorithms; accurate cross-modal alignment to ensure precise pixel correspondences across all modalities, facilitating effective multimodal data-level fusion; and annotated raw thermal and UV data generated through a novel reprojection algorithm, which remaps ground-truth labels onto original distorted thermal and UV images, significantly reducing manual labelling effort and enabling exploration of alternative fusion and alignment methods. Additionally, MM5 incorporates diverse and challenging scenarios, including reflective surfaces, hot and cold objects, and varying illumination conditions, to ensure that each sensor modality captures distinct and complementary cues. By providing both aligned and unaligned annotations and accommodating raw and preprocessed data, MM5 supports comprehensive research across multiple tiers of multimodal fusion. Researchers thus have flexibility in developing and validating methodologies within diverse processing workflows and fusion strategies. The remainder of this paper details the MM5 data acquisition framework, elaborates on sensor calibration and alignment, describes the dataset structure, and discusses potential application domains. Preliminary segmentation experiments using the Segformer-based CMX model further demonstrate the dataset's utility, highlighting the impact of different modality combinations, preprocessing approaches, and lighting variations on robust multimodal fusion.

3.2.1 Key Contributions

The primary contributions presented in this paper are:

1. **A comprehensive multimodal dataset** 3.9**:** MM5 [12] integrates RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR) imagery, addressing existing gaps in publicly available multimodal datasets and enabling extensive multimodal fusion research.

2. **A robust and reproducible data acquisition and processing pipeline:** The pipeline systematically addresses consistent ambient conditions 3.5.1, sensor calibration 3.7 and modality alignment 3.A.3 in a labelling and post-processing pipeline 3.8, significantly enhancing dataset usability and facilitating accurate multimodal analysis.

3. **MAR: Multimodal Annotation Remapping** 3.8.1**:** A novel algorithm that reprojects ground-truth annotations onto original distorted thermal and UV images, enabling flexible experimentation with both aligned and unaligned data without the overhead of fully manual annotation.

4. **DTMRE: Deterministic Thermal Multi-Resolution Encoding** 3.10.1**:** An algorithm to provide stable 24-bit colour representation and enhanced thermal resolution, particularly in regions of interest. DTMRE transforms raw thermal data into a visually informative format, facilitating better feature extraction in thermal modality-based tasks.

5. **ADMRE: Adaptive Depth Multi-Resolution Encoding** 3.10.2**:** A depth preprocessing technique that adaptively enhances depth resolution in areas exhibiting significant spatial changes or regions of interest, improving the utility of depth information in multimodal fusion.

3.3 Related Work

Multimodal fusion in computer vision has been studied in diverse contexts, from autonomous driving to medical imaging. A recent systematic review by Brenner *et al.* [1] categorises RGB-D-thermal fusion techniques into pixel-level, feature-level, and decision-level approaches. Early works focused on hand-crafted feature fusion, for example, combining colour and thermal gradients for improved human detection [13, 14]. With the rise of deep learning, end-to-end networks that learn joint representations of multiple modalities have become prevalent. Examples include architectures for RGB-D semantic segmentation that integrate depth as an additional input channel or through modality-specific sub-networks and RGB-thermal CNN models for pedestrian detection and person re-identification [13, 15]. For instance, recent RGB-IR action recognition frameworks leverage cross-modal distillation to improve gesture recognition in the dark [16]. Moreover, research has explored attention mechanisms, modality-specific encoding, and even modality hallucination or translation (e.g., predicting thermal from RGB) to handle missing or noisy inputs [15]. Recent advances have further enriched this landscape with innovative deep fusion strategies and generative approaches for multimodal data. Guan *et al.* [17] introduced an illumination-aware deep neural network to fuse visible and thermal streams, thereby boosting pedestrian detection performance. Ma *et al.* [18] presented FusionGAN, a generative adversarial network that effectively generates fused infrared-visible images, preserving complementary features from both modalities. In the context of RGB–thermal fusion for high-level vision tasks, Tang *et al.* have contributed a series of works: a semantic-aware real-time fusion network for infrared-visible image fusion [19], a progressive illumination-aware model (PIAFusion) for multi-scale fusion [20], and a framework that rethinks the role of image fusion in object detection pipelines via progressive semantic injection [21]. Meanwhile, the fusion of RGB and depth data has also seen important developments. Mosella-Montoro and Ruiz-Hidalgo [22] developed a 2D–3D geometric fusion network that integrates colour images with depth maps using multi-neighbourhood graph convolutions, demonstrating significant improvements in indoor scene classification. These recent studies consistently report that multimodal input yields superior performance over single modalities, particularly under challenging conditions such as low lighting, camouflage, or sensor noise.

Despite progress, the community has lacked datasets to objectively benchmark multimodal fusion methods beyond the RGB-D or RGB-thermal pair. Addressing the limitations identified in existing literature and enabling comprehensive multimodal fusion research by providing an extensive, well-structured dataset that includes raw and preprocessed data, as well as aligned and unaligned annotations, is the primary focus of our work.

3.3.1 Existing Multimodal Datasets

Until recently, most datasets were limited to two modalities, such as NYU Depth (RGB, Depth) [23] or KAIST Multispectral (RGB, Thermal) [24]. Here, we highlight efforts that combine three or more sensor streams. Palmero *et al.* (2016) introduced one of the first triple-modal datasets, the VAP Trimodal People Segmentation Dataset [13], featuring approximately 11.5k frames (5.7k labelled) of indoor scenes with synchronized RGB, Kinect depth, and thermal infrared footage. Although spatially calibrated and annotated with per-pixel human masks, the dataset is limited to three static scenes with few participants, underscoring the need for broader-scale trimodal datasets. More recently, Stippel *et al.* (2023) expanded significantly on this concept with the TRISTAR dataset [14]. TRISTAR includes 15,618 frames of tri-modal RGB, depth, and thermal streams across ten distinct indoor environments, along with semantic segmentation and temporal action detection annotations. For face anti-spoofing, Zhang *et al.* (2019) created the large-scale dataset CASIA-SURF capturing RGB, depth, and near-IR video streams of 21k sequences from 1,000 individuals [25]. This dataset provides annotations distinguishing real and spoofed faces, facilitating multimodal anti-spoofing research. In gesture recognition, the MGR-Dark dataset by Shi *et al.* (2024) comprises over 31k video clips of dynamic hand gestures recorded simultaneously in RGB, depth, and IR modalities under varying lighting conditions [16]. A dataset targeting autonomous driving scenarios is InfraParis, introduced by Franchi *et al.* (2024), it contains 7,301 street-view images with RGB, thermal IR, and depth modalities [26], alongside detailed semantic segmentation and bounding box annotations. Similarly, Baltaxe *et al.* (2023) presented a polarimetric dataset featuring synchronised polarimetric, RGB, and LiDAR imagery captured across diverse road conditions [27]. In robotics, comprehensive multi-sensor datasets such as the Multi-modal and Multi-scenario SLAM Dataset for Ground Robots (M2DGR) [28], incorporating RGB, thermal, and event-camera imagery, LiDAR scans, IMU measurements, and GPS information, have significantly advanced research in robot localisation and SLAM applications. This trend extends into specialised agricultural robotics, exemplified by datasets such as CitrusFarm by Teng *et al.* (2023) [29],

which combines RGB stereo, depth, monochrome, NIR, and thermal imagery in seven extensive sequences (1.3 TB total) captured using a custom multi-sensor rig, specifically designed for crop monitoring and robotic navigation tasks. Similarly, FieldSAFE (Kragh *et al.*, 2019) [30] provides approximately two hours of ROS-bagged RGB stereo (including 360° panoramic), thermal, LiDAR, and radar data, annotated for obstacle detection in farming environments. The Fruity dataset (Abdulsalam *et al.*, 2023) [31] further addresses precision agriculture with 11,065 annotated RGB, depth, and thermal images alongside pose data. Depth images are provided in 16-bit format, whereas thermal data are available in 8-bit. Conversely, Navarro et al. (2022) released a novel ground-truth multispectral image dataset of grape berries (grape Berries), including weight, anthocyanin, and Brix index measures, designed for machine learning applications [32]. It offers high-resolution 37-band VIS–NIR multispectral images for food quality assessment but lacks explicit annotations. For salient object detection, the VDT-2048 dataset (Song *et al.*, 2022) [33] presents 2,048 RGB, depth, and thermal image triplets. While it includes annotated ground truths, the thermal imagery is limited to 8-bit AGC-processed data and. We have summarised these key publicly available multimodal datasets, containing more than two imaging modalities, in Table 3.1. MM5 distinguishes itself from existing comparable datasets such as VDT-2048 not only through its inclusion of five spectral modalities and raw 16-bit data, but also by offering both aligned and unaligned annotations and a systematic set of eight controlled lighting conditions for RGB and three controlled settings for UV, consistently applied across all scenes. In contrast, VDT-2048 applies lighting variation sporadically and only to a subset of scenes, thus limiting its utility for comprehensive illumination-invariant fusion research.

While these existing multimodal datasets have demonstrated the advantages of integrating multiple imaging modalities, each exhibits certain constraints. Typically, these limitations include a restricted number of modalities, the absence of simultaneous provision of both aligned and unaligned annotated data, or a lack of raw sensor data, restricting research flexibility. The MM5 dataset introduced in Section 3.9 addresses these shortcomings by offering a comprehensive multimodal collection comprising RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR) imagery. MM5 uniquely supplies raw and preprocessed depth and thermal data alongside aligned and unaligned annotations, enabling extensive exploration across all multimodal fusion and preprocessing levels. Furthermore, the dataset includes specifically designed scenes containing real fruits, plastic fruit replicas, partially rotten produce, and other challenging elements like reflective surfaces, providing distinct sensor cues for effective feature extraction and fusion. Additionally, hot and cold objects introduce temperature contrasts, reflective objects challenge depth sensing, and varying illumination conditions further enrich the dataset, empowering researchers to investigate various fusion strategies and assess modality-specific challenges. Thus, MM5 is a robust foundation for advancing multimodal fusion research and methodology development, as detailed in Section 3.9.

3.3.2 Multi-Resolution Thermal Encoding

Thermal image conversion from 16-bit to 8-bit presents a trade-off between preserving radiometric fidelity and enhancing contrast. Conventional methods like CLAHE [38, 39] and similar adaptive histogram techniques [40, 41] yield high-contrast images by dynamically adjusting mappings per frame [42]. However, such dynamic mappings can lead to inconsistencies in pixel intensities that complicate object detection [43]. In scenarios where subtle thermal gradients are critical—for example, in defect or anomaly detection—overly adaptive gain control may remove essential temperature distinctions, motivating interest in static or semi-static mappings that better preserve underlying differences [44, 45]. Multi-resolution approaches seek to balance global dynamic range compression with local detail preservation. Multi-scale Retinex strategies [46] and wavelet or pyramid-based schemes [40] decompose images into base and detail layers for selective contrast adjustment, although they may still rely on histogram-based operations and exhibit scene-dependent behaviour [47, 48]. Recent work has also focused on enhancing colour consistency and edge preservation using learned networks to refine multi-scale Retinex outputs [46] and multi-scale guided-filter methods for contrast enhancement [47]. While purely linear mappings, though stable, lack the benefits of multi-resolution detail enhancement [44]. Additionally, deep learning modules have been developed to learn optimal tone-mapping functions that maintain consistent colour palettes and structures [45].

In summary, prior work demonstrates a clear division: static linear approaches reliably maintain temperature references but often suffer from poor contrast, while adaptive methods, whether histogram-based or multi-scale, dramatically improve detail at the cost of interframe consistency. Recent multi-resolution methods [46, 40, 47, 48] aim for a middle ground, retaining subtle distinctions in local intensities while compressing overall scene brightness. Nonetheless, a fully static colour mapping with multi-resolution temperature partitioning remains only partially

Table 3.1: Comparative summary of publicly available multimodal datasets with more than two imaging modalities or distinct wavelengths.

Dataset (Year)	Sensors / Specs	D Res	T Res	# Samp.	Domain	Modalities Present									Anno		Data Format						
						V	D	T	U	N	L	R	3	I	A	U	V	D	T	U	N	L	R
VAP Trimodal People Segmentation Dataset (2016) [13]	Kinect v2, FLIR A3	512x424	320x240	11.5k frames	Indoor segmentation	✓	✓	✓	–	–	–	–	–	–	✓	–	8	16	8	–	–	–	–
TRISTAR (2023) [14, 34]	Intel D435, FLIR Lepton	640x480	160x120	15.6k frames	Action segmentation	✓	✓	✓	–	–	–	–	–	–	✓	–	8	16	8	–	–	–	–
CASIA-SURF (2019) [25]	Intel SR300	640x480	–	21k clips	Face anti-spoofing	✓	✓	–	–	✓	–	–	–	–	✓	–	8	8	–	–	8	–	–
MGR-Dark (2024) [16]	Kinect v2 + IR	512x424	–	31k clips	Gesture recognition	✓	✓	–	–	✓	–	–	–	–	✓	–	8	16	–	–	8	–	–
InfraParis (2024) [26]	Stereo V, FLIR Tau2	–	640x512	7.3k images	Autonomous driving	✓	✓	✓	–	–	–	–	–	–	✓	–	8 S	disp	8	–	–	–	–
Polarimetic Imaging for Perception (2023) [27]	Polaris, DSLR, Velodyne	–	–	12.6k frames	Driving perception	✓	–	–	–	–	✓	–	–	–	✓	–	8	–	–	–	–	pts	–
M2DGR (2022) [28]	Multi-cam, FLIR Boson	–	640x512	36 sequences	Robotics SLAM	✓	–	✓	–	–	✓	–	–	✓	✓	–	8	–	8	–	–	pts	–
CitrusFarm (2023) [29]	Stereo V-D, FLIR T, NIR	–	640x512	7 sequences	Agriculture (robotics)	✓	✓	✓	–	✓	–	–	–	✓	✓	–	8 S	disp	8	–	8	–	–
FieldSAFE (2019) [30]	Multisense V, FLIR A65, Velodyne, Radar	–	640x512	2 hrs ROS data	Agriculture (obstacles)	✓	✓	✓	–	–	✓	✓	✓	✓	✓	–	8 S	disp	8	–	–	pts	rd
Fruity (2023) [31]	Stereo V-D, Thermal	640x480	640x480	11.0k images	Fruit picking	✓	✓	✓	–	–	–	–	–	–	✓	–	8	16	8	–	–	–	–
Grape Berries (2022) [32]	Lab (LED)	–	–	1.3k images	Food Quality	–	–	–	–	✓	–	–	–	–	–	–	–	–	–	–	16	–	–
KAIST (2018) [35]	Flea3, FLIR A655Sc, Velodyne	–	640x480	95k images	Driving	✓	–	✓	–	–	✓	–	–	–	✓	–	8 S	s	8	–	–	pts –	
WMCA (2020) [36, 37]	Basler acA1920/acA1921, Xenics Gobi-640, Intel D415	640x480	640x480	2904 sequences	PAD	✓	✓	✓	–	✓	–	–	–	–	✓	–	8	8	8	–	–	–	–
VDT-2048 (2022) [33]	Kinect v2, T640	512x424	640x480	2048 images	SOD	✓	✓	✓	–	–	–	–	–	–	✓	–	8	16	8	–	–	–	–
MM5 (2025) [12]	Azure Kinect, iRay Micro 640, Sony XC-EU50	1024x1024	640x512	324 scenes; 2592 images	SOD; Multimodal Fusion	✓	✓	✓	✓	✓	–	–	–	✓	✓	✓	8 S	16 Ss	16	8	16 S	–	–

D = Depth, T = Thermal, L = LiDAR, R = Radar, U = UV, N = NIR, 3 = 360-degree camera, I = IMU, Res = Resolution.
"Anno" = Annotations; "A" indicates for aligned data and "U" indicates for unaligned data.
“8” or “16” indicates bit-depth for that modality, and "S" indicates that stereo images are available; “disp” is disparity; “s” for D is stereo RGB available but no disparity; “pts” is point cloud; “rd” is range–Doppler. A dash (–) denotes absence. “✓” indicates presence.

explored. Our proposed DTMRE technique complements this area by employing a fixed, discrete set of colour gradients in conjunction with multi-resolution segment interpolation, effectively ensuring reliable per-temperature encoding without losing the critical detail required for accurate object detection and classification.

3.3.3 Multi-Resolution Depth Encoding

Time-of-flight (ToF) sensors provide high-precision depth maps for robust 3D reconstructions and object detection [49, 50, 51]. However, managing large 16-bit depth data requires efficient compression. Straightforward methods, such as mapping depth to a hue channel [50], 8-bit quantisation [52], or using lossless formats like PNG, often compromise local detail and hinder tasks like boundary delineation [53]. Advanced approaches have been developed to address these limitations. For instance, Wilson's RVL codec exploits runs of similar depth values to reduce storage while retaining detail [54]. Region-based techniques divide scenes into planar or smoothly varying areas for piecewise encoding [49, 55], although they add overhead for segmentation. Specialised colourisation methods and learning-based frameworks further leverage correlations between RGB and depth data [52, 56]. For downstream tasks, the influential HHA encoding transforms raw depth into three channels (horizontal disparity, height above ground, and the angle with the surface normal) to augment RGB-based networks [57, 58]. However, HHA's reliance on accurate ground-plane estimation and its significant preprocessing cost in cluttered scenes pose challenges [59, 56]. Adaptive strategies, such as saliency-driven segmentation, aim to allocate higher resolution to regions of interest [60, 55], yet they often require complex preprocessing and may struggle with irregular shapes [49, 58].

Our proposed ADMRE technique differs fundamentally from existing methods by leveraging Kernel Density Estimation (KDE) on the raw 16-bit depth distribution to detect peaks and by adaptively compressing these peak regions with a finer resolution while assigning coarser resolutions to Out-of-Focus (OOF) or low-variation depth ranges. Unlike segmentation-based approaches that may require prior object detection or planar fitting, we directly derive compression rules from data-driven density estimates. In addition, the design accommodates a two-channel (24-bit) encoding with up to 980 discrete depth steps, leaving the third channel for optional surface normals or other features. This end-to-end pipeline ensures minimal detail loss in critical regions, leading to improved performance in subsequent detection and segmentation tasks compared to conventional uniform quantisation.

3.4 Multimodal Hardware System

The data capture setup integrates multiple sensing modalities to enable comprehensive scene analysis. The system comprises two Microsoft Azure Kinect sensors, a Sony XC-EU50/CE ultraviolet (UV) camera, and an iRay Micro 640 long-wave infrared (LWIR) thermal module. The thermal module is securely mounted using a custom-engineered 3D-printed frame to ensure precise spatial alignment and minimise cross-modal misalignment. This design maintains the geometric consistency of all sensors, reducing errors caused by displacement and ensuring stable multimodal image acquisition. The calibrated setup facilitates reproducible data capture across varying environmental conditions, enhancing the integrity of multispectral data fusion.

The sensor array configuration includes:

- **2× Azure Kinect RGB-D sensors**, capturing both visible spectrum images and near-infrared (NIR) depth information at 850 nm.
- **1× Sony XC-EU50/CE UV camera**, designed for imaging within the 300–420 nm spectral range, enabling ultraviolet feature extraction.
- **1× iRay Micro 640 LWIR module**, operating within the 8–14 µm infrared spectrum for thermal imaging and heat signature analysis.

This multimodal setup forms the foundation for the MM5 dataset by ensuring accurate and consistent data capture across diverse spectral domains. By carefully aligning and calibrating each sensor, we establish a reliable framework for generating high-quality, multimodal training data, facilitating advancements in cross-spectral learning and sensor fusion research, as depicted in Figure 3.1.

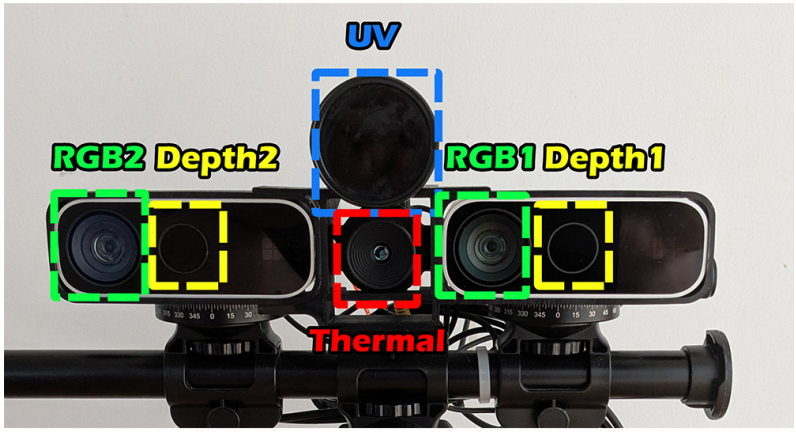


Figure 3.1: Multimodal sensor array comprising 2× RGB, 2× Depth + NIR (850 nm), 1× LWIR (8–14 µm), and 1× UV (300–420 nm).

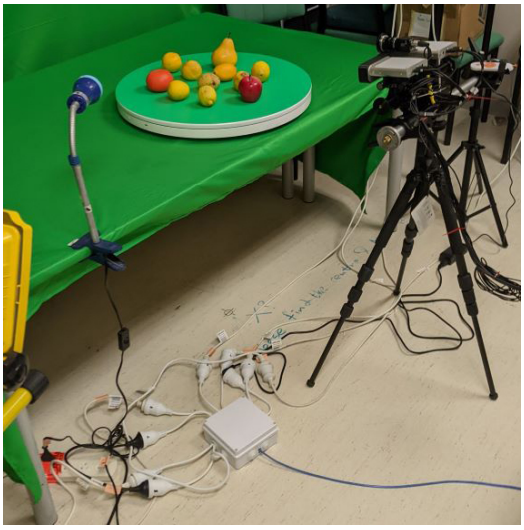


Figure 3.2: Capturing setup in a controlled laboratory environment.

3.5 Capturing Setup

We set up the data capture system in a controlled laboratory environment to ensure consistent and reproducible acquisition conditions across all modalities. Ambient lighting remained constant throughout all recording sessions to minimise external illumination variability. We utilised a green cloth backdrop for certain scenes to simplify background segmentation and enable subsequent replacement, thus facilitating data augmentation through background variation. However, the backdrop introduced unintended reflections of infrared (IR) illumination emitted by the Time-of-Flight (ToF) depth sensor, resulting in gaps within the depth data. These artefacts were deliberately retained in the dataset, presenting a realistic challenge. Furthermore, to maintain consistent exposure levels across captures, the RGB cameras were operated with fixed exposure settings, preventing automatic adjustments that could introduce inconsistencies in brightness and contrast. This setup allows for controlled acquisition of overexposed and underexposed scenes, ensuring a diverse dataset that reflects real-world lighting conditions. To further guarantee stability and precise alignment during data capture, the sensor array was securely mounted into a custom-designed, rigid 3D-printed frame, which was firmly affixed to a tripod (Figure 3.1). We performed calibration of the thermal, UV, and right RGB cameras relative to the left RGB camera, thereby establishing precise spatial correspondence across all imaging modalities. The complete data acquisition setup is shown in Figure 3.2.

3.5.1 Lighting

The MM5 dataset employs a varied lighting configuration designed to closely simulate a broad spectrum of real-world illumination conditions. To ensure comprehensive coverage of realistic scenarios, we systematically utilised eight distinct light sources during the data acquisition process: LED 1, LED 2, Desk lamp (60W), UV 365nm, Halogen Floodlight, Desk lamp (purple LED), UV 365nm Spot, and Halogen Spot. We defined nine different lighting settings to capture multimodal data. For each setting, we captured images from 8 RGB channels, 3 UV channels, one thermal channel, one depth channel, and one infrared channel. The lighting settings were as follows:

- Setting 1: Dimmed room light (with dim UV illumination)
- Setting 2: Sidelight from the right
- Setting 3: Full illumination (optimal lighting)
- Setting 4: Backlight combined with thermal illumination
- Setting 5: Overexposure
- Setting 6: Dimmed light from the left (using a purple LED)
- Setting 7: Low UV/halogen spotlight
- Setting 8: UV 365 nm illumination
- Setting 9: UV overexposure

This varied lighting configuration ensures that the dataset captures a broad spectrum of illumination conditions, enhancing its utility to evaluate multimodal fusion techniques and robust performance in diverse environments.

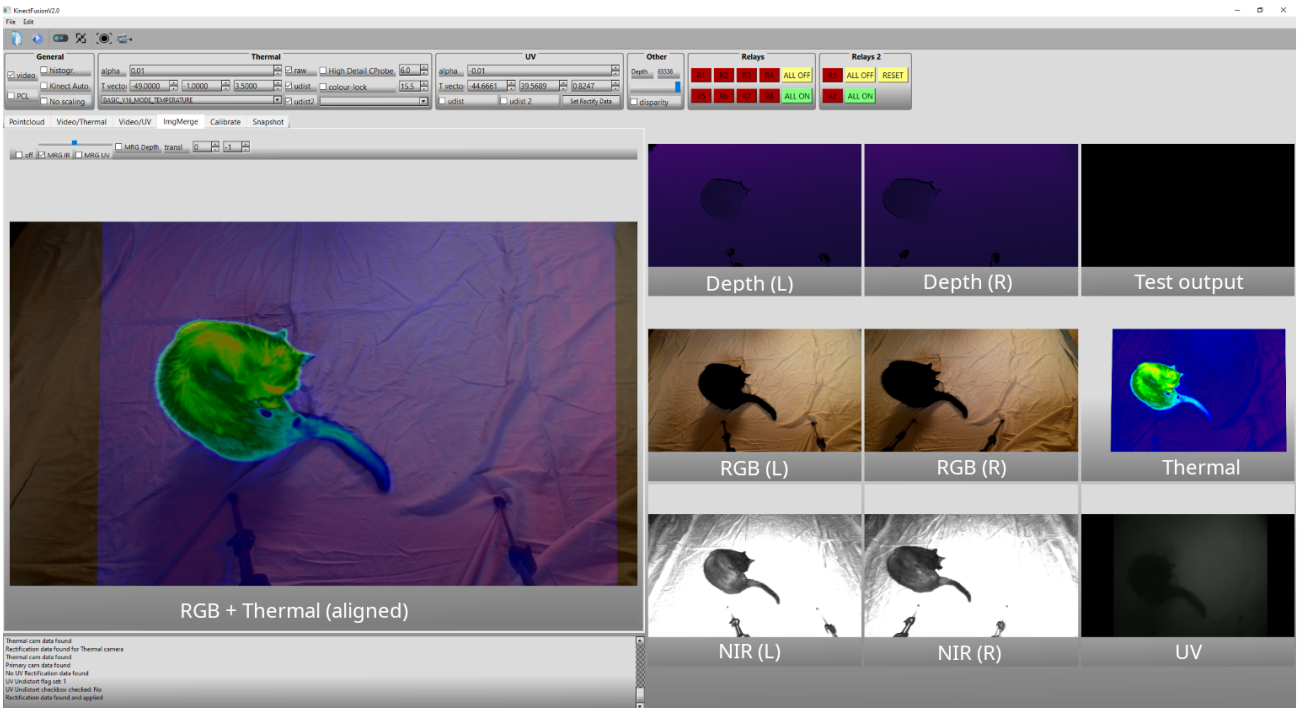


Figure 3.3: Screenshot of the capturing software with real-time overlay and rectified thermal data showing a cat. The panel on the left shows the thermal data aligned with the left RGB data. The nine panels on the right, starting from top left to bottom right, show depth L+R, test output, RGB L+R, thermal, NIR L+R and UV

3.6 Capturing Software

To develop a high-performance software solution for multimodal data acquisition, we implemented the system in C++ due to its efficiency, low latency, and robust hardware integration capabilities. The software interfaces with the Kinect SDK and the thermal imager SDK while simultaneously integrating the UV camera video stream and capturing metadata for each acquisition. This setup enables synchronised image capture across all modalities, continuous video stream recording, and incorporates a stereo calibration algorithm [61] conveniently integrated with OpenCV [62]. This calibration algorithm facilitates the estimation of the translation vector, rotation matrix, and distortion coefficients. We store the raw images and corresponding camera parameters, enabling alignment computations in a dedicated post-processing stage, which ensures flexibility for refining and optimising alignment. Additionally, to enhance automation and dataset diversity, the system incorporates a relay array for dynamically controlling scene illumination and a motorised turntable that rotates objects by 120 degrees, facilitating the acquisition of three distinct viewpoints per scene. Figure 3.3 shows the capture and real-time alignment that allows for the testing and fine-tuning of the obtained camera parameters.

3.7 Camera Calibration And Registration

For successful multimodal data level fusion using multiple modalities, it is crucial to acquire the data from these modalities correctly aligned. This can pose a challenge, as the sensors used for each modality have different fields of view (FOV), resolutions, and sensing capabilities. To facilitate data-level fusion, the system was calibrated by determining the intrinsic (pinhole camera model parameter matrix) and extrinsic (estimation of the relative sensor poses) parameters of each camera, which can then be used to align the data. Based on the pinhole camera model, this calibration has been simplified using a stereo calibration process[61], which can be applied using these and similar modalities. This method has been implemented in numerous studies in different ways, as summarised by Brenner *et*

al. [1]. Figure 3.5 shows the pattern matching using stereo calibration. However, uneven and fluctuating heat can complicate the calibration of the thermal camera. To achieve a more uniform heat distribution, the backside of the calibration board was covered with copper plates, as shown in Figure 3.4, and a heating mat was placed over them. This approach helped to stabilise the temperature during calibration. Additionally, since the UV camera can detect wavelengths extending into the lower bounds of the visible spectrum (up to 420 nm), we removed the UV filter lens during the alignment process. This allowed us to use the captured grayscale image for alignment without requiring a dedicated setup for the UV modality. Furthermore, because stereo calibration requires uniform image resolutions, we applied lens distortion correction, scaling, and padding to the thermal and UV images prior to calibration. Figure 3.5 shows an example set of RGB and thermal calibration images, including the calibration pattern matches generated by the stereo calibration algorithm.

Figure 3.4: Calibration board backside with partially applied copper plates.

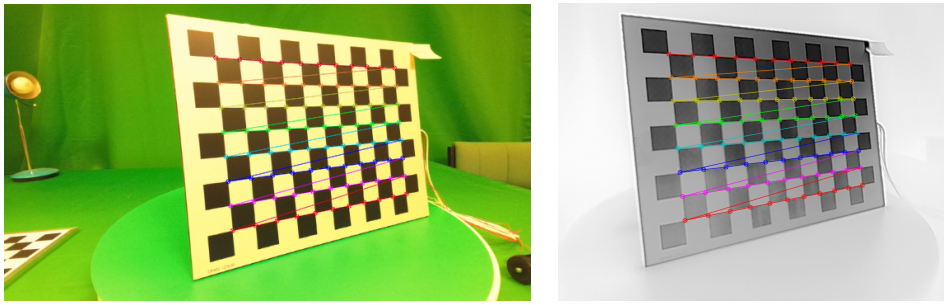


Figure 3.5: Calibration images of RGB (left) and thermal (right) captured at approximately 30 cm from the sensor. The overlaid lines illustrate the pattern recognition process of the stereo calibration.

After calibrating the intrinsic and extrinsic parameters of the RGB, thermal, and UV cameras, the thermal and UV images are aligned to the RGB coordinate system through a projection transformation. However, this alignment is optimised for a single reference plane in the scene; objects that lie substantially nearer or farther than this plane exhibit misalignment due to parallax effects. Consequently, the current approach ensures robust alignment in the primary plane of interest, with gradually increasing deviations as objects move away from that plane. Because the Kinect SDK automatically aligns the depth data to the RGB camera, we can directly utilise it without further modification. However, the intensity image, representing the 850 nm near-infrared (NIR) reflectance, is not a standard output of the Kinect SDK and requires additional processing. There are two primary approaches to obtaining this image. The first method extracts intensity values using the depth alignment process, ensuring direct correspondence with the depth map but limiting intensity information to pixels with valid depth data. The second method employs a synthetic flat depth image to bypass this limitation, allowing the retrieval of a complete intensity image that captures the full 850 nm NIR reflectance. However, as this image retains the field of view (FOV) of the depth sensor rather than the RGB camera, post-processing is required to correct FOV discrepancies. Additionally, the alignment process performed by the Kinect SDK operates discretely across depth intervals due to the differing sensor FOVs, employing a closed-source algorithm. Although this process helps mitigate geometric distortions, it also introduces occlusion artefacts such as depth shadows and missing regions. To address these limitations, depth information from the stereo setup can be fused, reducing alignment inconsistencies and improving overall depth map completeness. Additionally, we include the calibration image sets and corresponding calibration data as part of our dataset, ensuring transparency and reproducibility of the calibration process while enabling researchers to validate or develop alternative calibration methods.

3.8 Labelling and Post-Processing Pipeline

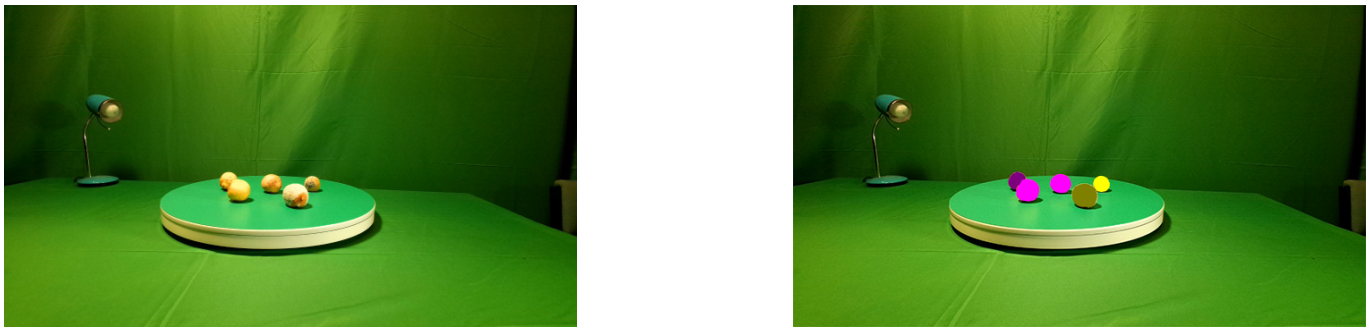
The labelling and post-processing pipeline is critical in preparing the dataset for multimodal analysis. This process ensures that annotations are consistently applied across different imaging modalities while maintaining spatial alignment between RGB, depth, thermal, and ultraviolet (UV) data. Given the inherent differences in sensor properties, including variations in field of view (FOV), resolution, and spectral characteristics, an efficient annotation workflow is necessary to achieve high-quality labelled data suitable for downstream tasks such as object detection, segmentation, and classification. For this, we employed a structured annotation workflow to facilitate accurate labelling, initially focusing on the RGB images. Once annotated, these labels were reprojected onto the thermal and UV images

using transformation matrices derived from the camera calibration. This method ensures that the annotations are accurate and consistent across modalities, enabling cross-modal learning and sensor fusion. Following the automated labelling process, a series of post-processing steps are applied to refine the annotations and improve alignment. These steps include depth-based adjustments and annotation corrections to compensate for the different viewing angle.A detailed overview of the entire pipeline is shown in Figure 3.14 and sections describing the annotation framework in Label Studio, the label export process and the post-processing techniques applied to ensure high-quality multimodal alignment can be found in 3.A while algorithm Alg. 1 below summarises the key steps in the multimodal image alignment process.

Algorithm 1 Multimodal Image Alignment

1: *Load camera calibration parameters* (including intrinsic and extrinsic matrices).
2: *Read raw images* from all modalities (RGB, Thermal, UV, Depth, and IR).
3: *Histogram Equalisation:* For Thermal and IR images, apply histogram equalisation to generate an 8-bit representation.
4: *Rectify images* to correct lens distortion in Thermal and UV data using the calibration parameters.
5: *Align images* to the RGB coordinate system by applying the appropriate transformation matrices.
6: *Apply perspective correction* to compensate for differences in field-of-view (FOV) across sensors.
7: *Reprojection for Thermal and UV using MAR:* Apply inverse distortion corrections and reversed alignment transformations using the inverse camera matrix to remap the RGB labels onto the original, distorted Thermal and UV images.
8: *Save outputs* in both full-resolution and cropped formats.

3.8.1 MAR: Multimodal Annotation Remapping Algorithm



(a) RGB image without labels.

(b) RGB image with labels.

Figure 3.6: RGB images showing colourised label overlay. The left image is without labels, and the right image includes manually created labels. The three pink-toned lemons on the left are good, while the two yellow-toned lemons indicate mold.

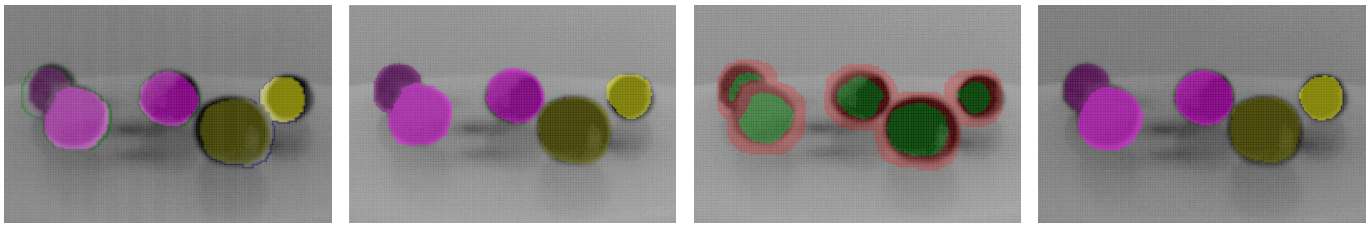
We generate labels for the unprocessed ultraviolet (UV) and thermal images to facilitate the exploration of fusion methods beyond early data-level fusion, which typically requires extensive calibration and can introduce latency in real-time processing. Our labelling approach, MAR (Multimodal Annotation Remapping), partially eliminates the labour-intensive task of manually labelling these modalities by reversing the same forward transformation that aligns thermal and UV images with the RGB labels. In essence, the RGB annotations, accurately aligned with the RGB images as shown in Figure3.6, are remapped onto the thermal and UV images(in their rectified state) and then distorted by reapplying lens distortion using the inverse of the original alignment transformation. This novel reverse mapping technique, to our knowledge, has not been reported previously in the literature, where multimodal fusion methods typically focus on early data-level fusion. In most cases, this approach yields accurate results, and even when minor discrepancies occur, it places the correct labels near their intended locations. This is particularly valuable given the inherent ambiguity in UV and thermal data, where insufficient contrast often makes it challenging to distinguish between classes clearly. Figure3.7 illustrates the target labels after processing, demonstrating how inverse camera

matrices and transformation parameters transfer the labels from the RGB domain to the target modalities. In summary, MAR proceeds in four key steps: (1) inverse mapping (remapping annotations onto raw thermal/UV images), (2) re-distortion (re-applying lens distortion models), (3) depth-based refinement (adjusting labels based on average depth to account for FOV differences), and (4) edge-guided region growing (refining labels using a random walker [63] approach guided by Canny edges [64]). Although depth-based refinement, demonstrated in Figure 3.8 using the Random Walker algorithm [63, 65], is a key component of our approach to maximise alignment accuracy, it can be omitted when depth information is unavailable or insufficient. In such cases, MAR still performs reliably by relying on the inverse transformation and re-distortion steps, followed by edge-guided refinement. Additional pseudocode and implementation details are provided in Section 3.D, where we sequentially describe each step: starting with inverse mapping and re-distortion (Section 3.D.2), and concluding with depth-based correction (Section 3.D.3) and edge-guided random walker refinement (Section 3.D.4). A flowchart summarising the overall sequence of steps is provided in the appendix (Figure 3.17). The MAR algorithm primarily serves as an automated label transfer and initialisation tool to reduce annotation effort. The remapped labels are intended for manual refinement, supported by subsequent edge-guided and machine learning-based segmentation techniques, to produce high-quality ground truth. Our qualitative visualisations in Figures 3.7 and 3.8 demonstrate that MAR provides spatially consistent and accurate label placements across modalities, substantially accelerating the annotation process and improving consistency compared to manual labelling from scratch.

(a) Thermal image without labels.

(b) Thermal image with labels.

Figure 3.7: Thermal images illustrating reprojection results, shown without labels (left) and with calculated labels (right).



(a) Remapped and re-distorted labels

(b) Labels after depth correction.

(c) Thermal image with label seed areas in green and growth areas in red for Random Walker.

(d) Labels after Random Walker.

Figure 3.8: MAR refinement process. (a) Initial remapped and re-distorted labels. (b) Labels after depth-based correction. (c) The calculated seed (green) and growth (red) areas. (d) The final, refined labels after applying the Random Walker [63, 65] optimisation.

To quantitatively evaluate the accuracy of MAR-generated annotations, we compared them against the final manually corrected labels using mean Intersection over Union (IoU) and pixel accuracy metrics per class. The results, summarised in Table 3.10 in the appendix, demonstrate that MAR-generated labels closely match the manual corrections,

achieving mean IoU and accuracy values generally between 80% and 90% for most classes. These findings confirm that MAR provides a reliable automatic initialisation that substantially reduces manual annotation effort and facilitates the generation of high-quality labels.
Overall, this novel approach, which combines inverse geometric transformation, depth-based FOV correction, and edge-guided region growing, constitutes a robust solution for efficiently transferring annotations from RGB to thermal and UV domains.

3.8.2 Final Image Generation and Storage

Once the images have been aligned, they undergo additional processing to optimise their usability for downstream applications. Specifically, the aligned images are cropped to a standardised size within the fully overlapping region across modalities. To achieve this, we compute the centroid of the RGB label and adjust the cropping window so that this point is as central as possible while ensuring that the target resolution is entirely contained within the overlapping area. This approach preserves critical target information while retaining background variation, thereby supporting robust analysis in subsequent tasks.

3.9 MM5 Dataset

This section provides an overview of the MM5 dataset [12]. Each subsection below describes a specific component of the dataset, outlining the characteristics and challenges associated with each modality. Together, these elements form a robust resource for research in multimodal sensor fusion and advanced computer vision applications.

3.9.1 Structure

The raw data is organised into separate folders for each camera. In our stereo setup, the left and right camera images are suffixed with `_0` and `_1`, respectively. Each image file follows a standardised naming convention that includes the sequence number, settings ID, timestamp, and modality as a postfix. The raw data folder structure is as follows:

- `DEPTH_0`
- `DEPTH_1`
- `IR_0`
- `IR_1`
- `LWIR`
- `META`
- `RGB_0`
- `RGB_1`
- `UV`
- `ANNO_V`
- `ANNO_T`
- `ANNO_U`

An example filename for an RGB image with light setting 5 is: `1_5_20240716_130310_143_rgb.png`.
The captured raw data and the transformation outputs provided by the Kinect SDK are available for the depth and IR modalities. These files are differentiated by a postfix: `_raw` for raw data and `_tr` for transformed data. Both depth and IR images are stored as 16-bit single-channel images. For the thermal (LWIR) modality, multiple representations are provided:

- Raw 16-bit images (`_lwir16`)
- 24-bit fixed colour encoded images (`_lwir`)
- 8-bit normalised grayscale images (`_lwir8dyn`)

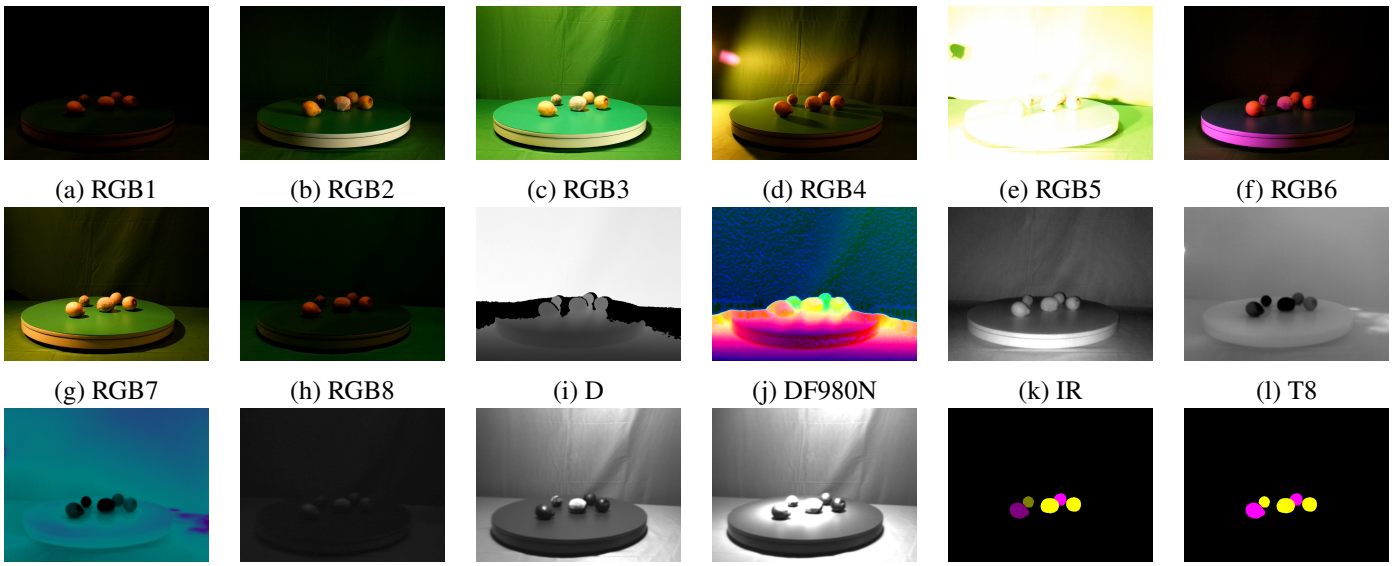
The encoded LWIR images are included for convenience, as they can be derived from the raw data.
The aligned and cropped dataset is generated by selecting a subset from the raw data, since the raw data includes additional unlabeled images, and renaming these selected images sequentially, starting from 1, to ensure consistent filenames across all modalities captured simultaneously. The processed data are organised into the following folders:

- `ANNO_CLASS`
- `ANNO_INST`
- `ANNO_VIS_CLASS`
- `ANNO_VIS_INST`
- `D`
- `D_Focus`
- `D_Focus960N`
- `D16`
- `I`
- `I16`
- `META`
- `RGB1`
- `RGB2`
- `RGB3`
- `RGB4`
- `RGB5`

- `RGB6`
- `RGB7`
- `RGB8`
- `T8`
- `T16`
- `T24`
- `U1`
- `U8`
- `U9`

It is worth noting that while the RGB and UV folders are named according to their corresponding light setting, the thermal, IR, and depth folders are distinguished by their encoding type. The folders prefixed with `ANNO_VIS_` contain colour-coded class and object instance labels for visualisation purposes, whereas the actual annotations are stored in `ANNO_CLASS` and `ANNO_INST`. To avoid unnecessary duplication, we refrained from creating redundant copies of modalities common across multiple lighting configurations. For example, depth, thermal, and infrared data, which remain constant across RGB captures under different illumination settings (`RGB1`, `RGB2`, etc.), are provided only once. Researchers wishing to train on multiple or all lighting configurations will thus need to pair the shared depth, thermal, or IR data explicitly with each RGB setting. This pairing can be readily implemented through custom data loader scripts tailored to specific network architectures, or by restructuring the dataset into the desired format as needed.

3.9.2 Images



(a) RGB1 (b) RGB2 (c) RGB3 (d) RGB4 (e) RGB5 (f) RGB6

(g) RGB7 (h) RGB8 (i) D (j) DF980N (k) IR (l) T8

(m) T24 (n) U1 (o) U8 (p) U9 (q) AI (r) AC

Figure 3.9: Set of images from the dataset.

Figure 3.9 displays the same scene captured under eight different light settings, as described in Subsection 3.5.1. The figure includes eight RGB images (RGB1 through RGB8) that illustrate the effects of varying illumination alongside a depth image (D), a processed depth image (DF980N) and an infrared image (IR). It also shows two thermal encodings (T8 and T24) and three UV images (U1, U8, and U9). Additionally, the annotation images, for object instance (AI) and class (AC) labels, are provided to demonstrate the corresponding ground truth. These images offer a comprehensive view of the scene and underscore the diverse conditions captured in the MM5 dataset.

3.9.3 Thermal Raw Data

The thermal raw data are stored as 16-bit unsigned integers, representing temperature measurements in a scaled format. To convert the raw data into degrees Celsius, the raw data are first scaled to Kelvin by dividing by 64 and then converted to Celsius by subtracting 273.15 as per the formula below:

$$T_{\text{Celsius}} = \frac{T_{\text{raw}}}{64} - 273.15, \tag{3.1}$$

where T_{raw} is the 16-bit raw thermal value. This formula directly interprets the sensor data in standard temperature units.

3.9.4 Thermal 8-bit Data

This version of the thermal image is generated using a multi-stage adaptive tone-mapping algorithm designed to produce a visually optimised 8-bit output. The process begins by dynamically suppressing intensity outliers based on histogram percentiles. Subsequently, the remaining pixel values are stretched to maximise the dynamic range, followed by an adaptive gamma correction that non-linearly enhances detail in regions corresponding to the lower end of the frame's temperature range. The final enhanced 16-bit data is then normalised to an 8-bit representation, yielding an image with pronounced thermal contrast suitable for qualitative analysis.

3.9.5 Thermal 24-bit Data

The thermal 24-bit data in the MM5 dataset is produced by applying our novel DTMRE algorithm for static colour mapping to the raw temperature values. This mapping utilises a predefined gradient of distinct colours, each corresponding to specific temperature intervals. For further details, please refer to Section 3.10.1.

3.9.6 Depth Raw Data

The raw depth data are stored in millimetres. Each 16-bit integer value represents the distance from the sensor to objects in the scene, providing high-precision measurements suitable for further processing and analysis.

3.9.7 Depth 8-bit Data

For convenience, we provide an 8-bit normalised depth image without inpainting. A zero-initialised 8-bit array is first prepared to store the final normalised depth values. Next, all non-zero pixels from the original 16-bit data are scaled into the 0–255 range using a min-max normalisation, thus preserving the relative distribution of valid depth measurements. The zero or invalid pixels remain untouched, retaining their values in the 8-bit representation. This process yields a visually consistent depth image highlighting contrasts among valid regions.

3.9.8 Depth Focused 8-bit Data

This variant of the depth image is produced by applying our novel ADMRE algorithm for the adaptive compression strategy detailed in Section 3.10.2, resulting in an 8-bit representation that preserves fine detail where depth variation is most pronounced. Non-essential regions are compressed at lower resolution, reducing noise and enhancing focus on critical structures. Consequently, the final 8-bit output provides a compact yet detailed view of salient depth information.

3.9.9 Depth Focused 24-bit Data

The same ADMRE 3.10.2 algorithm as for the 8-bit focused data is used, but with an additional step that packs the data in a 24-bit format. The first two channels store the quantised depth values, while the third channel encodes the computed surface normals, allowing a visual representation of both geometry and spatial orientation. This approach offers a more complete scene depiction, combining depth-focused compression with local angular detail for downstream tasks.

3.9.10 Meta Data

During image capture, relevant metadata is recorded and stored in several files for subsequent analysis and annotation. The following files are generated:

- `label_mapping (.json/.csv)`: Provides a complete mapping from class labels to IDs, where each ID corresponds to a pixel value in the annotated images.
- `label_instances.json`: Contains mappings between images, classes, instances, label names, and Label Studio IDs.
- `classes.txt`: A list of all classes, sorted by their corresponding IDs.
- `dataset_meta.csv`: A comprehensive list of all dataset files with their associated labels and challenges.

- `filename (.json/.csv)`: Stores all metadata associated with each image file, including the IMU data from the Kinect sensor.

Semantically, this metadata file encapsulates rich contextual information for each captured sample. The `category` and `subcategory` fields provide hierarchical classification, with the first indicating a general class (e.g., *Fruit*) and the latter a specific type (e.g., *Mandarin*). The `challenge` field documents the conditions under which the sample was captured (such as *real*, *rotten* or *good*), which can be used to evaluate the robustness of the algorithm under varied conditions. The `green_screen` flag denotes a green screen during capture, facilitating background segmentation. In addition, the `master_imu` and `subordinate_imu` sections record inertial measurement data, providing information on the orientation and motion of the sensor at the time of capture. The `sequence` number serves as a unique identifier for each capture and is incorporated into the filenames of the raw data. In the aligned and cropped dataset, this identifier is retained in the metadata file as a reference to the original capture sequence, thereby preserving the connection between the raw and processed files even after renaming. Collectively, these elements define the semantics of the dataset, enabling comprehensive multimodal analysis and supporting advanced sensor fusion techniques.

3.9.11 Labels

In the MM5 dataset, semantic annotations are provided at the pixel level using two distinct labelling schemes: class labelling and object instance labelling. In the class labelling scheme, each pixel value directly corresponds to a specific class defined in the file `classes.txt`, ensuring consistency with the predefined category list. In contrast, object instance labelling assigns a unique pixel value to each object instance within a given class, thereby enabling the differentiation of multiple objects of the same class.

Figures 3.9r and 3.9q illustrate these two approaches. The former, AC (Annotation Class), displays the class labels, while the latter, AI (Annotation Instance), shows the object instance labels. This dual annotation strategy conforms to standard practices in semantic and instance segmentation that are exemplified by widely used datasets such as PASCAL VOC [66], MS COCO [67], and NYU [23], as well as by segmentation methods such as Mask R-CNN [68]. This approach facilitates both class-level analysis and object-level detection.

3.9.12 Classes

Currently, the dataset comprises 14 top-level classes and a total of 32 labelled classes across 324 scenes, as detailed in Table 3.2. Ongoing efforts will continue to expand the dataset by incorporating additional scenes, videos, and annotated classes in future releases.

Table 3.2: Label Distribution Table

Class	Total Frames	Subclass
Lemon	80	Good, Bad, Fake, Half
Mirror	29	
Bowl	26	
Mandarin	57	Good, Bad, Fake, Half, Peel
Kettle	8	
Cup	61	Hot, Cold
Onion Red	21	
Onion	21	
Grapes Green	42	Good, Bad, Fake
Grapes Blue	32	Good, Bad, Fake
Apple	32	Good, Fake
Apple Green	56	Good, Bad, Fake
Pear	30	Good, Bad
Carrot	30	Good, Fake

The chart in Figure 3.10 illustrates the total class distribution, highlighting the relative frequency of each class variant within the total frames that encompass the class.

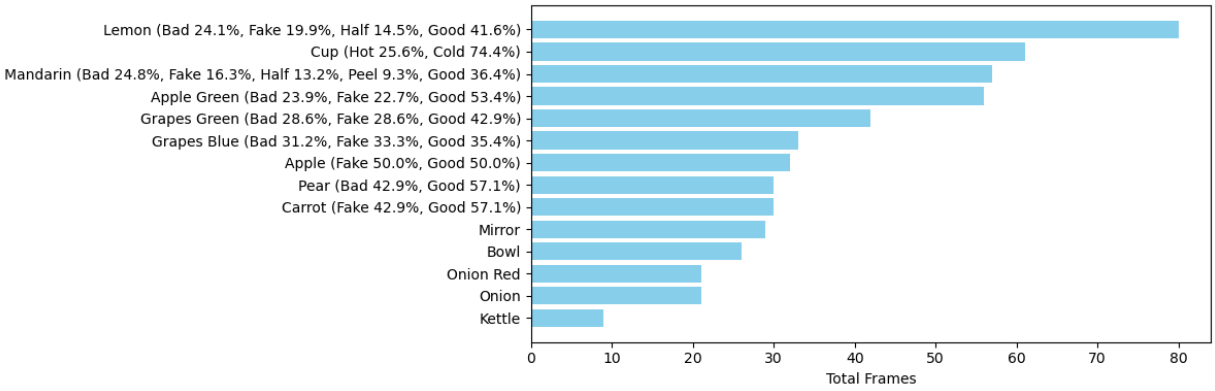


Figure 3.10: Distribution of unique label occurrences. Percentages shown in parentheses represent the frequency of subclass occurrences (e.g., good, bad, fake) among all class occurrences. Due to mixed-scene acquisitions, subclasses may co-occur together within a single frame.

3.9.13 Calibration Data

We provide calibration data stored in YAML files as part of the dataset. These files contain essential parameters including intrinsic camera matrices (CM1 and CM2), distortion coefficients (D1 and D2), the rotation matrix (R) and translation vector (T), as well as the essential (E) and fundamental (F) matrices. In addition, the files include the rectification transforms (R1 and R2), the projection matrices (P1 and P2), and the Q matrix for reprojection. This data is provided in the following files:

- `def_stereocalib_THERM.yml`
- `def_stereocalib_UV.yml`
- `def_thermalcam_ori.yml`
- `def_uvcam_ori.yml`
- `def_stereocalib_cam.yml`

The YAML calibration files `def_stereocalib_THERM.yml` and `def_stereocalib_UV.yml` contain the calibration results for the thermal and UV cameras relative to the left RGB camera. The file `def_stereocalib_cam.yml` provides the calibration data to align the right and left Kinect sensors, enabling accurate stereo image alignment and depth data fusion. The calibration files with the `_ori` suffix (`def_thermalcam_ori.yml` and `def_uvcam_ori.yml`) represent the data obtained using the original resolutions of the cameras to calculate lens distortion. The images used to obtain the calibration data are provided in the calibration subfolder. This comprehensive calibration information facilitates reproducibility and enables other researchers to apply or develop their own calibration algorithms.

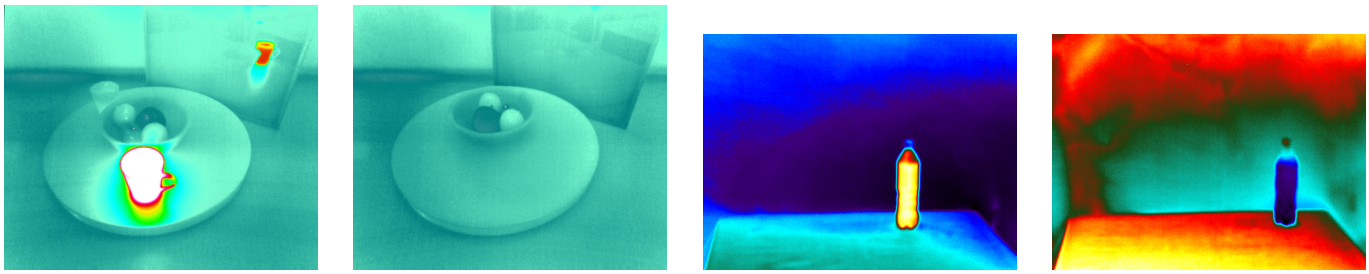
3.10 Data Pre-Processing

In the MM5 dataset, we provide the raw data for thermal and depth modalities, both stored in a single channel 16-bit format. Preprocessing techniques are applied to enhance these modalities for subsequent visual analysis. Depth data are refined using our novel range-based detail enhancement algorithm, ADMRE, and the raw thermal data is converted into a consistent 24-bit colour image via a novel static colour mapping approach, DTMRE, enhancing thermal resolution for the temperature ranges of most interest.

3.10.1 Processing Thermal Data With DTMRE

The proposed 24-bit thermal data representation in the MM5 dataset applies a *static* colour mapping to raw temperature values, thereby addressing long-standing issues of inconsistent contrast that arise in dynamically adjusted schemes [38, 39]. Specifically, a predefined gradient is segmented into multiple intervals and populated via linear interpolation at each step. This gradient, spanning 2455 distinct colours, transitions through a series of perceptually distinct hues (e.g. black to teal, teal to purple), ensuring that temperature distinctions remain visually salient, resulting in a refined palette that conveys subtle temperature cues. The raw thermal readings are first translated into degrees

Celsius by applying Equation 3.1. The temperature scale is then divided into segments of varying resolution: a fine-grained *core* interval (for example, 14°C to 30°C), surrounded by broader segments for higher or lower temperatures (e.g. 30°C to 40°C, 40°C to 100°C, and 0°C to 14°C). A piecewise linear function determines the appropriate index for a given temperature, ensuring that out-of-bounds values are clamped. By returning the colour at the computed index, this scheme produces a stable 24-bit representation that faithfully captures pixel-wise thermal variation. Table 3.3 outlines the primary colour transitions used by this method.



(a) MM5_541_1_lwir (b) MM5_539_1_lwir (c) VDT2048_26 (d) VDT2048_18

Figure 3.11: Comparison of 24-bit thermal image representations. Images (a) and (b) show MM5's DTMRE method, while images (c) and (d) depict images from the VDT2048 dataset processed with automatic gain control (AGC) and a colour scheme. Images (a) and (c) show hot objects, whereas images (b) and (d) represent cold objects.

This static mapping method overcomes the shortcomings of Automatic Gain Control (AGC) algorithms, which convert raw 16-bit data (0–65,535) into 8-bit images (0–255), compressing the data and reducing detail. Although AGC algorithms enhance contrast and brightness to emphasise contextual details [69], their dynamic adjustments can lead to inconsistent representations between frames. In contrast, our novel DTMRE method preserves environmental features regardless of object temperature. Figure 3.11 shows two sets of thermal images: images `a` and `c` depict a hot object, while images `b` and `d` represent cold objects. In the dynamic range images taken from the VDT2048 dataset [33] (images `c` and `d`), noticeable variations in the appearance of the background and table are observed despite no actual changes, whereas our DTMRE processing maintains a stable visual representation. Even with a cup of boiling water in the frame, the minute temperature differences of the fruit in the bowl are visible. This consistency is crucial for the reliable processing of thermal data in form of visual information. It should be noted that slight variations in the thermal output may occur over time and during non-uniformity correction (NUC) procedures for uncooled thermal cameras, potentially introducing minor discrepancies in the measured values and, consequently, in the visual appearance of the images. The consistency provided by the DTMRE encoding is crucial when detecting minute temperature differences, such as those between rotten and good fruits. The complete pseudocode for the DTMRE algorithm is provided in

Table 3.3: Defined primary colour transitions used in our DTMRE demonstration.

Colour Transition	**Index Range**	**Colour Transition**	**Index Range**	**Colour Transition**	**Index Range**
Black to Purple	0 – 130	Aquamarine to Cyan	851 – 977	Dark Orange to Dark Red	1683 – 1782
Purple to Navy	131 – 180	Cyan to Green	978 – 1232	Dark Red to Red	1783 – 1837
Navy to Deep Blue	181 – 207	Green to Lime Green	1233 – 1424	Red to Magenta	1838 – 2092
Deep Blue to Blue	208 – 323	Lime Green to Gold	1425 – 1487	Magenta to Rose	2093 – 2200
Blue to Grey	324 – 523	Gold to Yellow	1488 – 1527	Rose to Light Pink	2201 – 2285
Grey to Teal	524 – 723	Yellow to Orange	1528 – 1617	Light Pink to White	2286 – 2435
Teal to Aquamarine	724 – 850	Orange to Dark Orange	1618 – 1682		

Section 3.B.

Evaluation

To evaluate the performance of the T8 (normalised 8-bit thermal image 3.9.4) and T24 (24-bit DTMRE processed thermal image) representations, we process the full 324 scenes from the MM5 dataset. The training and evaluation images are defined in the files list_train_f.txt and list_eval_f.txt, respectively. We utilised a CMX model [70], a SegFormer-based segmentation network, and only used low-light (RGB1) images as supplementary input. This design ensures that thermal data remains the dominant source of information for segmentation. We have trained each model

from scratch for 500 epochs on the MiT-B0 backbone, and the final segmentation metrics are summarised in Table 3.4. Detailed class-level results can be found in Table 3.7.

To provide a comprehensive benchmark, we compare our methods against several established contrast enhancement techniques. We implemented Contrast Limited Adaptive Histogram Equalisation (CLAHE) using the standard OpenCV library function, which performs localised histogram equalisation to improve detail [71]. A clahe_clip_limit of 30.0 and a clahe_tile_grid_size of (24, 24) were used. For Plateau Histogram Equalisation (PHE), we implemented the algorithm by clipping the image histogram at a plateau level of 10 before applying equalisation, a method known to control noise amplification [72]. Finally, we included Multi-Scale Retinex (MSR), implemented using a standard three-scale configuration with Gaussian surround functions to enhance dynamic range and colour constancy. The sigma scales for the three Gaussian paths were set to 15, 80, and 250, respectively [73].

The quantitative evaluation, summarised in Table 3.4 and detailed in Table 3.7, demonstrates that the proposed 24-bit DTMRE thermal representation (T24) consistently outperforms both the normalised 8-bit thermal image (T8) and the established contrast enhancement methods PHE, CLAHE, and MSR. Specifically, T24 achieves the highest mean Intersection over Union (79.08%) and mean pixel accuracy (86.67%), surpassing MSR and T8, which attain a mean IoU of 72.67% and 72.29% and a mean pixel accuracy of 82.93% and 81.94% respectively. Among the traditional methods, Multi-Scale Retinex (MSR) performs best with a mean IoU of 72.67%, followed by CLAHE (69.12%) and PHE (60.52%). Notably, T24 provides marked improvements in challenging classes such as *Mandarin Peel* and *Pear Bad*, indicating its superior ability to preserve thermal details critical for segmentation accuracy. These results highlight the effectiveness of the DTMRE encoding in enhancing thermal image quality for robust multimodal fusion, particularly under low-light conditions where thermal cues are essential.

To facilitate equitable comparison between the algorithms, we also report the Mean Rank for each method [74]. Within each class, algorithms are ranked according to their performance, with rank 1 assigned to the best performing method, and ties receiving an average rank. The mean rank of each algorithm is then calculated as the average of its class-wise ranks, offering an interpretable, class-balanced summary of comparative performance across the full class set.

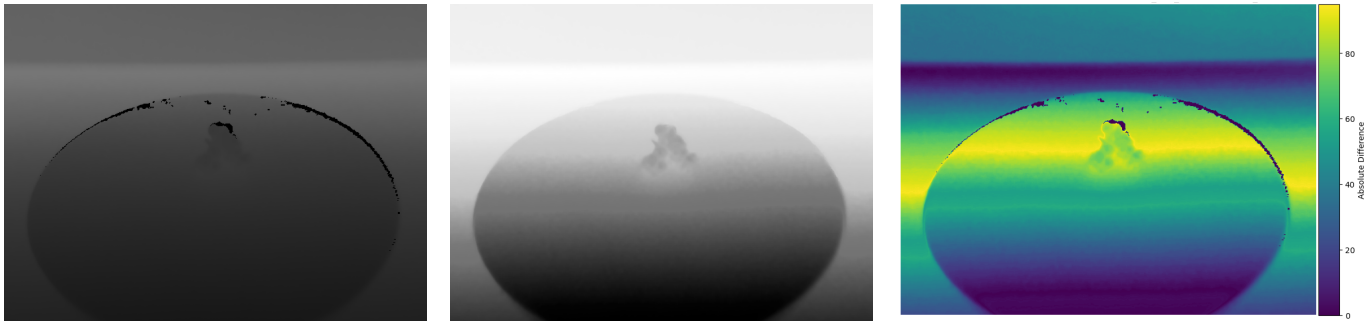
Table 3.4: Performance of Thermal Image Preprocessing. A comparison of our proposed DTMRE method (`T24`) against established baseline algorithms (`PHE`, `CLAHE`, `MSR`) and our normalised (`T8`) image. Metrics include Intersection over Union (IoU), Pixel Accuracy (Acc), and Mean Rank, with IoU and Acc reported as percentages (%).

	PHE	CLAHE	MSR	T8	T24
Mean IoU	60.52	69.12	72.67	72.29	**79.08**
Freq IoU	98.62	98.76	99.02	99.05	**99.23**
Mean Pixel Acc	73.34	80.76	82.93	81.94	**86.67**
Pixel Acc	99.18	99.29	99.44	99.45	**99.58**
Mean Rank	4.42	3.55	2.80	2.73	**1.50**

3.10.2 Processing Depth Data With ADMRE

In the MM5 dataset, depth information is captured using a Time-of-Flight (ToF) sensor, which intrinsically provides high-fidelity range measurements with millimetre resolution. Over a distance of approximately 3 metres, these sensors generate up to 3000 discrete depth values, increasing further for longer ranges. Although such fine granularity is beneficial for applications such as object detection and segmentation [49, 52], directly storing and processing 16-bit depth data can be computationally expensive and memory-intensive. A common practice is to convert depth values into an 8-bit format for efficiency and visualisation. However, this uniform quantisation compresses essential structural details, often degrading the accuracy of downstream vision tasks [53]. To address these limitations, we propose a novel adaptive method that allocates a higher effective depth resolution to regions that exhibit significant depth variations while assigning reduced resolution to homogeneous areas. This concept of directing compression resources to salient parts of the scene aligns with other regions of interest strategies [60]. However, our approach uses data-driven kernel density estimation (KDE) [75, 76] to identify the most critical depth intervals. For this a KDE of the of the density distribution is computed, revealing prominent peaks corresponding to depth ranges in which notable variation occurs. For platforms where real-time performance is critical and full KDE evaluation is too slow, an alternative approximation can be employed by computing a histogram with a reduced number of bins and then smoothing it using a Gaussian

filter. This approach yields a similar density estimate with significantly lower computational cost, enabling faster peak detection.
Although producing an 8-bit grayscale representation is often convenient for compatibility with standard image pipelines, it remains limited by the 255 discrete values available for encoding depth. To mitigate this constraint, we extended the algorithm to support a 24-bit depth representation in which the red and green channels jointly store up to 980 distinct depth values, leaving the blue channel available for encoding additional data. This design is partly inspired by colourisation methods [50], though our specific use of multiple channels ensures finer control over resolution. Crucially, the increased capacity allows surface normals—computed directly from the final depth map—to be saved as pixel intensities, thus facilitating a compact yet interpretable encoding of scene geometry. Storing normals has been shown to improve depth-based object detection performance, as it highlights critical shape and orientation cues [57].



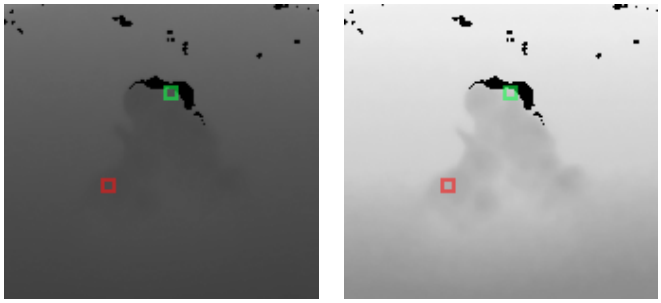
(a) (b) (c)

Figure 3.12: (a) Normalised 8-bit image, (b) focused 8-bit image, and (c) Pixel-wise absolute difference between ADMRE processed and normalised image.

In general, this hierarchical focus strategy enables an efficient and informative compression of ToF depth data. By adaptively assigning resolution according to local depth variability rather than applying a uniform quantisation scheme, the method preserves key structural information paramount for object detection, scene segmentation, and other high-level tasks. The two-tier design (8-bit vs 24-bit output) allows users to trade-off between universal compatibility and high-resolution detail, illustrating the framework's flexibility. As such, the proposed pipeline addresses the known trade-off between resource constraints and preservation of local detail in ToF depth data [54, 58], offering an efficient, data-driven alternative to conventional compression and encoding methods.
The complete pseudocode for the ADMRE algorithm is provided in Section 3.C.

Evaluation

Figure 3.12 compares normalised 3.12a and 8-bit ADMRE-processed 3.12b depth representations of an example scene. A detailed close-up of an object of interest (grapes) in this scene is shown in Figures 3.13a and 3.13b, highlighting two specific regions (marked in green and red) selected for pixel intensity measurement. In addition to these processed representations, intensity values were recorded from equalised and raw depth data as well. These measurements are summarised in Table 3.5. Furthermore, Figure 3.12c visualises the absolute pixel-value differences between the normalised and ADMRE-processed images, emphasising pronounced deviations within the detected peak region.
The results demonstrate a value difference of 33 in the raw data, corresponding to a distance of 33 mm. In the ADMRE processed image, there are 30 values for this region, whereas in the normalised image, we obtained only 14 values. Additionally, the histogram's equalisation yielded a mere seven value difference. These effects amplify with the distances present in a frame, as a more pronounced initial value range results in a greater compression of values during the normalisation or equalisation process. We evaluated the effectiveness of our approach using CMX with a Segformer-B0 backbone, a state-of-the-art vision transformer framework for segmentation, trained from scratch on 324 scenes of the MM5 dataset for 250 epochs. The training and evaluation images are defined in the files `list_train_f.txt` and `list_eval_f.txt`, respectively. Table 3.6 summarises the mean IoU per method (%) for the three variants: standard depth (**D**), depth focus (**DF**), and depth focus with 980 discrete levels plus normals (**DF980N**) as



(a) (b)

Figure 3.13: (a) Zoomed-in normalised 8-bit image, and (b) zoomed-in focused 8-bit image of grapes.

Table 3.5: Comparison of ADMRE processed (P), normalised (N), equalized (E), and raw (R) intensity data in the indicated green (G) and red (R) area of the images in Figure 3.12(c) and (d).

	G	**R**	**Difference**	**Resolution**
P	187	157	30	1.1 mm
N	96	82	14	2.4 mm
E	96	89	7	4.7 mm
R	738	704	33	1 mm

well as HHA encoded depth. The full class-level results can be found in the appendix in Table 3.9. Since the network only has access to visual and geometric cues, discriminating among defective (*bad*), plastic (*fake*), and (*good*) produce classes poses a significant challenge due to the similarity of the objects in shape and colour.

Table 3.6: Overall mean IoU (%) for each depth encoding method combined with RGB3, along with summary metrics and estimated processing time per image (ms) for 640×480 resolution. DF980N achieves both superior segmentation accuracy and significantly lower computational cost compared to HHA.

Metric	**D** (%)	**DF** (%)	**HHA** (%)	**DF980N** (%)
Mean IoU	70.66	71.97	72.02	**76.33**
Freq IoU	99.14	99.17	99.16	**99.27**
Mean Pixel Acc	81.32	80.53	81.59	**84.37**
Pixel Acc	99.48	99.51	99.50	**99.57**
Mean Rank	2.95	2.72	2.70	**1.62**
processing time (ms)	0	23	250-500	25

Despite these inherent difficulties, the results in Table 3.6 indicate that **DF** and **DF980N** outperform the baseline **D** in terms of mean IoU, frequency-weighted IoU, and pixel-level accuracy. Importantly, **DF** and **DF980N** can be computed in approximately 25ms per image, whereas HHA encoding is substantially slower, requiring over 500ms per 640×480 image with the original method and still around 260ms even with a recent optimised implementation [77]. We attribute the observed segmentation improvements to the selective preservation of fine-grained depth details in salient image regions. In contrast, normalised 8-bit depth often fails to capture subtle curvature and spatial variation crucial for distinguishing nuanced object classes [53]. By dedicating higher resolution to areas with strong depth gradients while simultaneously allowing coarser encoding in low-variation zones, our approach helps minimise quantisation artefacts like contour banding or edge distortion that degrade downstream performance [55, 60]. These findings align with previous research on region-based compression and saliency-driven encoding [49], confirming that refined geometric features can significantly enhance scene segmentation [57]. More importantly, they demonstrate that curated depth encoding can serve as a decisive factor in differentiating between visually similar categories, thus validating the key design goals of our proposed method.

3.11 Challenges and Future Work

Despite the significant advantages offered by the MM5 dataset, several challenges remain, presenting opportunities for further research and development.

- **Multimodal Sensor Calibration and Alignment:** Although comprehensive calibration procedures have been implemented to align RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR) sensors, misalignments can still occur due to lens distortions, differences in fields of view (FoV), and parallax effects. More robust calibration and alignment methods, potentially aided by deep learning, could further mitigate alignment errors, compensate for parallax-induced discrepancies, and ensure consistent feature fusion across modalities.

- **Reflective Surfaces and Missing Data:** Scenes containing mirrors, metallic objects, or liquids can result in regions of missing or invalid depth data due to reflections or sensor interference. Although these scenarios are realistic and highlight the importance of robust data preprocessing, the resulting missing or noisy depth data remains a significant challenge. Future studies could focus on advanced hole-filling strategies or deep learning-based inpainting methods that retain cues about reflective surfaces.

- **Thermal Consistency Under Varying Conditions:** The 24-bit static thermal mapping in MM5 counters the dynamic nature of AGC algorithms, but uncooled cameras can still exhibit measurement fluctuations and drifts, especially during extended operation or pronounced environmental changes. Non-uniformity correction (NUC) processes may also introduce subtle shifts or discontinuities over time. Future work could investigate techniques to minimise these effects, enabling stable thermal representations and ensuring reliable pixel-wise temperature measurements under variable conditions.

- **Large-Scale Data and Real-Time Fusion:** Although MM5 includes diverse objects, lighting conditions, and sensor modalities, further expanding the dataset with extensive indoor and outdoor scenes would enhance its benchmarking capabilities. Many real-time applications, such as robotics and AR/VR, require efficient on-device multimodal data fusion. Future studies could focus on developing fusion pipelines optimised for real-time processing in resource-constrained environments.

- **Annotated Multimodal Datasets with Complex Tasks:** The current version of MM5 supplies aligned and unaligned annotations suited to segmentation, object detection, and classification tasks. However, more advanced challenges remain. For instance, spatio-temporal action recognition would require temporal annotations and sequences of frames across modalities. At the same time, 3D reconstruction and material classification would require correspondingly richer labels detailing structure and surface properties. Extending the dataset with these additional annotations and establishing benchmarks that span multiple tasks would enable researchers to investigate modality-specific effects more thoroughly and advance the development of robust fusion techniques that leverage the complementary strengths of each sensor modality.

By addressing these technical and methodological challenges, future investigations can expand the value of MM5 beyond its current scope. Although replicating the sensor configuration may be non-trivial, the dataset and its accompanying resources are designed to foster new calibration, preprocessing, and fusion techniques that exploit the complementary advantages of multiple spectral channels. These efforts, in turn, will continue to drive progress in multimodal computer vision research and applications.

3.12 Conclusion

We have introduced the MM5 dataset, a comprehensive multimodal imaging resource that integrates RGB, depth, thermal, ultraviolet, and near-infrared modalities. Although variations in fields of view can introduce parallax effects that limit perfect alignment outside the central overlap region, thorough sensor calibration and alignment mitigate these discrepancies for most of the crucial, labelled areas. Additionally, MM5 offers raw 16-bit data for depth and thermal measurements, allowing researchers to investigate sophisticated preprocessing and data fusion techniques that leverage unique spectral information across modalities. To further enhance usability, we developed novel algorithms for thermal colour encoding, depth focus compression, and multimodal annotation remapping, addressing critical challenges such as consistent temperature representation, adaptive depth resolution, and flexible automated label generation. Our preliminary experiments with a transformer-based segmentation network illustrate the potential for improved performance when leveraging MM5's data and preprocessing techniques. By addressing these technical and methodological challenges, future investigations can expand the value of MM5 beyond its current scope. Although replicating the sensor configuration may not be trivial, the dataset and its accompanying resources are designed to foster new calibration, preprocessing, and fusion techniques that exploit the complementary advantages of multiple spectral channels. These efforts, in turn, will continue to drive progress in multimodal computer vision research and applications.

Data Availability

The MM5 dataset introduced in this paper is publicly available under the following Figshare links: (1) Raw data `https://figshare.com/ndownloader/files/55555451`, (2) Aligned and cropped data `https://figshare.com/ndownloader/files/55555457`, (3) Label Studio annotations `https://figshare.com/ndownloader/files/55555424`, and (4) Calibration images `https://figshare.com/ndownloader/files/55555421`. Additional resources, code examples, and updates are available via the project repository: `https://github.com/martinbrennernz/MM5-Dataset`.
If you use this dataset in your research, please cite both this paper and the dataset DOI [12], for example:

> M. Brenner, N. Reyes, T. Susnjak, and A. Barczak (2025). MM5: Multimodal Image Dataset. Figshare. Dataset. `https://doi.org/10.6084/m9.figshare.28722164`

3.A Labelling Process

3.A.1 Labelling Process Overview

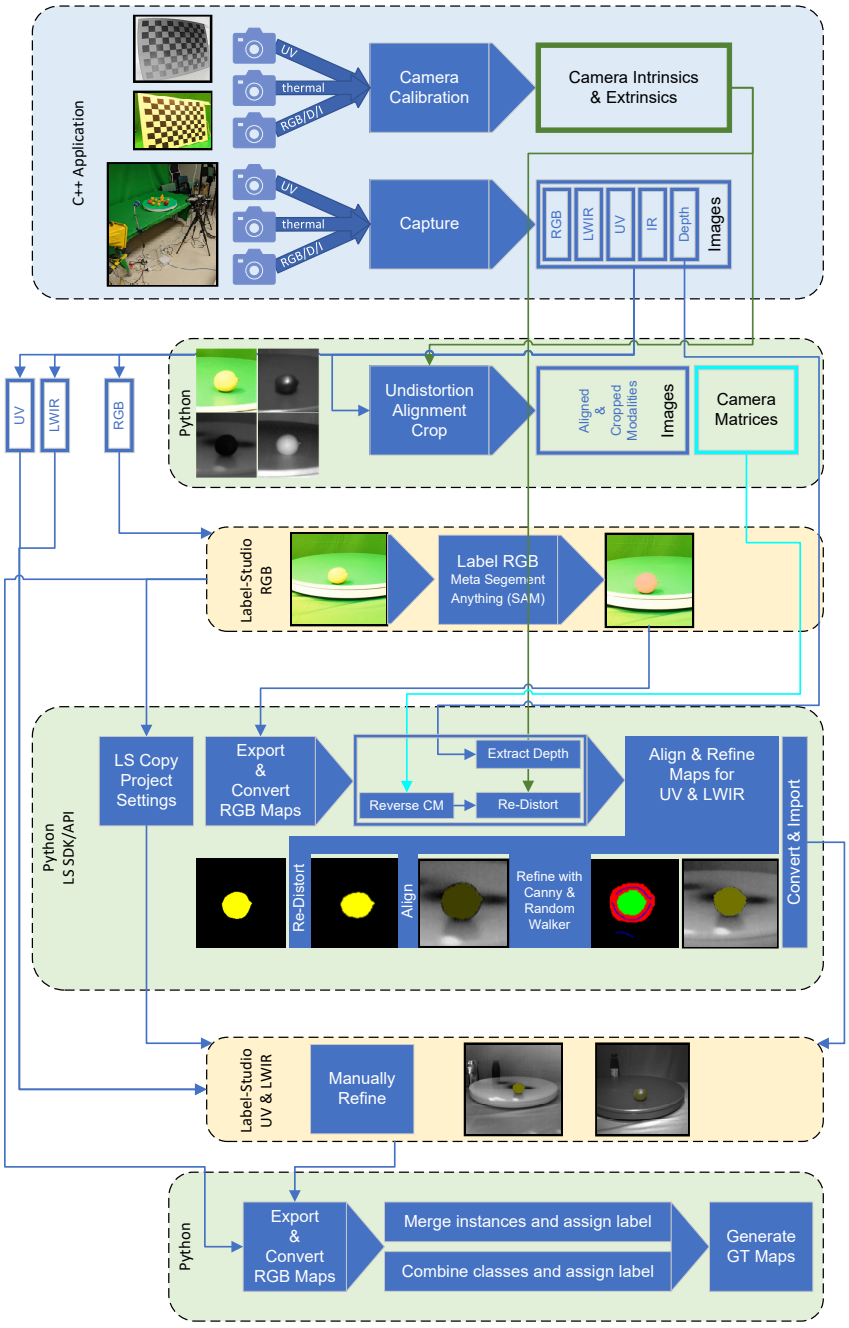


Figure 3.14: Labelling process pipeline

3.A.2 Labelling and Post-Processing: Label Studio

To facilitate the annotation process, we utilised *Label Studio*, an open-source data labelling platform designed for multimodal annotation tasks [78]. Label Studio provides a flexible web-based interface that supports various annotation types, including image segmentation, classification, and object detection. The platform's ability to handle custom labelling workflows made it well suited for our multimodal dataset, where labels initially created for the RGB images are later reprojected onto the thermal and UV modalities. One of the primary advantages of using Label Studio is its compatibility with deep learning-assisted annotation. To enhance the efficiency and accuracy of the annotation process, we integrated Meta's *Segment Anything Model* (SAM) [79] to help label RGB images. SAM is a powerful image segmentation model that enables automatic region selection, significantly reducing the manual effort required to annotate complex scenes. By leveraging SAM within Label Studio, we streamlined the annotation pipeline, improving consistency and minimising human-induced errors. In a first step, RGB images were set up for annotation, and later projects were configured for the thermal and UV modalities by copying the RGB configuration. When the RGB labels are reprojected onto the thermal and UV images using transformation matrices obtained from the multimodal calibration process, the resulting labels are created in those target projects. Label Studio provided several key benefits in the annotation and dataset management workflow for MM5: it offered a web-based interface that enabled efficient collaborative annotation across multiple modalities, facilitating streamlined data labelling; it supported deep learning models by integrating with segmentation models such as SAM to pre-label data, significantly reducing manual annotation effort; it allowed multimodal compatibility for handling RGB, thermal, and UV images within a unified framework, ensuring consistency of annotation across different sensor outputs; it featured export options in JSON and COCO formats, facilitating seamless integration into the downstream processing pipeline; and it offered a Python SDK [80] integration that supported automated project setup, replicate configurations, and extracted metadata, with export/import functionality allowing calculated UV and thermal labels to be re-imported for iterative refinement. By incorporating Label Studio into our annotation workflow, we established a scalable and efficient framework to generate a high-quality labelled dataset across multiple imaging modalities, ensuring that the labelled data remain consistent, well organised, and aligned for subsequent multimodal analysis.

3.A.3 Multimodal Image Alignment and Processing

Aligning multimodal images involves rectifying and registering data from different sensors to ensure spatial consistency across all modalities. This section describes the pipeline used to generate accurately aligned images from the RGB, depth, thermal, and ultraviolet (UV) captures by applying camera calibration parameters and the re-projection of labels onto the original, distorted thermal and UV modalities. The process begins with loading the camera calibration parameters, where the intrinsic and extrinsic parameters of each sensor are retrieved from the calibration files. Next, the RGB annotations on class and instance levels are exported from Label Studio. Subsequently, the captured RGB, thermal, UV, depth, and infrared (IR) images are read into the processing pipeline. Distortion correction is then applied to thermal and UV images using the camera's intrinsic parameters, after which thermal and UV images are aligned to the RGB coordinate system using projection transformation matrices. Additional postprocessing transformations are applied to correct minor misalignments caused by field-of-view differences based on the depth of labelled objects. Finally, the pipeline produces cropped and full-resolution aligned outputs for further processing. A flow diagram summarising these steps is shown in Figure 3.14, providing an overview of the entire multimodal image alignment process. The following subsections detail the methods used for Image Rectification and Alignment, the Processing Pipeline Algorithm, and the MAR: Multimodal Annotation Remapping Algorithm.

Image Rectification and Alignment

Since each sensor possesses different fields of view (FOVs), resolutions, and intrinsic distortions, a calibration step is required to ensure proper alignment. The intrinsic parameters (focal length and optical centre) define the pinhole camera model for each sensor, while the extrinsic parameters describe the transformations relating each camera. Thermal and UV images are first rectified to correct lens distortions, then projected into the RGB coordinate space using transformation matrices derived from calibration using OpenCV's [62] stereoRectify. At a high level, this process involves loading the relevant calibration parameters and performing rectification. Then we align the rectified thermal and UV image to the RGB frame, where a rotation adjustment is applied. A final refinement step adjusts

positioning. This approach ensures that thermal and UV modalities are spatially consistent with the RGB reference, enabling accurate downstream processing and annotation consistency across the modalities.

3.B DTMRE Implementation

3.B.1 Detailed Evaluation Results

Table 3.7: Per-class mean Intersection over Union (IoU, %) for thermal preprocessing methods including Plateau Histogram Equalization (PHE), Contrast Limited Adaptive Histogram Equalization (CLAHE), Multi-Scale Retinex (MSR), MM5 normalised 8-bit thermal images (T8), and the proposed 24-bit DTMRE processed thermal images (T24), evaluated in combination with RGB1 dim light images. In addition to per-class IoU values, the table also reports mean pixel accuracy and the mean rank of each method across all classes, providing a comprehensive comparison of segmentation performance and ranking consistency over the full MM5 evaluation dataset.

	PHE	CLAHE	MSR	T8	T24
Background	99.61	99.61	99.71	99.71	**99.75**
Lemon	61.05	60.76	61.22	60.51	**72.06**
Lemon Bad	43.61	45.24	50.83	44.77	**63.89**
Lemon Fake	69.01	74.37	78.01	80.92	**84.65**
Mirror	97.89	95.55	98.07	98.21	**98.63**
Bowl	84.80	87.55	89.20	**90.42**	89.10
Mandarin	72.77	74.55	75.96	75.58	**76.74**
Mandarin Bad	37.64	38.07	39.62	32.29	**44.04**
Mandarin Fake	84.08	**84.90**	81.32	84.85	84.02
Kettle	88.97	89.59	89.07	92.39	**95.29**
Lemon Half	51.70	51.53	56.44	52.55	**70.91**
Mandarin Half	64.97	69.64	68.02	61.54	**70.13**
Mandarin Peel	50.70	57.93	44.87	50.11	**67.66**
Cup Hot	92.14	92.20	93.69	93.23	**93.84**
Onion Red	79.86	78.45	81.16	86.21	**89.33**
Onion	90.56	88.83	90.01	92.69	**93.66**
Grapes Green	55.89	72.38	55.95	56.63	**69.17**
Grapes Green Bad	30.66	56.70	81.35	82.18	**82.33**
Grapes Green Fake	39.21	39.97	53.65	48.94	**70.05**
Grapes Blue Fake	28.35	44.23	68.44	41.72	**83.19**
Grapes Blue	18.15	48.87	50.93	13.98	**65.44**
Grapes Blue Bad	62.80	80.75	85.66	87.58	**87.69**
Apple	25.35	70.07	71.09	89.62	**90.08**
Apple Fake	66.53	72.91	75.04	**90.36**	89.96
Apple Green	39.14	66.63	80.77	**89.62**	82.12
Apple Green Bad	33.87	55.90	61.96	**67.57**	57.87
Apple Green Fake	71.89	74.68	80.97	77.58	**83.34**
Cup Cold	87.35	85.47	81.23	**87.20**	85.91
Pear	34.88	42.83	49.30	**58.78**	53.68
Pear Bad	3.41	32.05	51.42	47.39	**56.33**
Carrot	84.47	89.75	88.91	88.39	**90.07**
Carrot Fake	85.26	89.70	**91.57**	89.63	89.62
Mean IoU	60.52	69.12	72.67	72.29	**79.08**
Freq IoU	98.62	98.76	99.02	99.05	**99.23**
Mean Pixel Acc	73.34	80.76	82.93	81.94	**86.67**
Pixel Acc	99.18	99.29	99.44	99.45	**99.58**
Mean Rank	4.42	3.55	2.80	2.73	**1.50**

3.B.2 DTMRE Algorithm

Algorithm 2 DTMRE: Deterministic Thermal Multi-Resolution Encoding

Inputs: **data_array** – 16-bit thermal raw data
Output: **data_array** – 24-bit processed thermal data with 3 colour channels

function ConvertThermalData(data_array) ▷ *data_array* is 16-bit thermal sensor output
  Ensure *data_array* is of type `np.uint16`
  Convert raw values to Celsius: $\textit{tempCelsiusArray} \leftarrow \frac{\textit{data\_array}}{64.0} - 273.15$
  (Optional) Clamp *tempCelsiusArray* to an application-specific temperature range (e.g. -50 to 150), noting that the choice of bounds may vary depending on the intended use case.
  Flatten *tempCelsiusArray* to obtain a 1D list
  Map each temperature value t to RGB via GetTemperatureColor(t)
  Reshape the resulting list back to $(\textit{height}, \textit{width}, 3)$
  Return the final 24-bit RGB image
end function
function GetTemperatureColor(tempCelsius) ▷ Maps temperature in °C to an RGB triplet
  $f_{0\max} \leftarrow 3,\ f_{0\text{res}} \leftarrow 0.01,\ f_{1\text{idx}} \leftarrow 400,\ f_{1\min} \leftarrow 14,\ f_{1\max} \leftarrow 30,\ f_{1\text{res}} \leftarrow 0.005$[1]
  $f_{2\max} \leftarrow 40,\ f_{2\text{res}} \leftarrow 0.02,\ f_{3\max} \leftarrow 100,\ f_{3\text{res}} \leftarrow 0.06,\ f_{\text{minres}} \leftarrow 0.1,\ \textit{index} \leftarrow 0$[1]
  if $f_{1\min} \le \textit{tempCelsius} < f_{1\max}$ **then**
    $\textit{index} \leftarrow f_{1\text{idx}} + \left\lfloor \frac{\textit{tempCelsius} - f_{1\min}}{f_{1\text{res}}} \right\rfloor$
  else if $f_{1\max} \le \textit{tempCelsius} \le f_{2\max}$ **then**
    $\textit{rangeInF1} \leftarrow \left\lfloor \frac{(f_{1\max} - f_{1\min})}{f_{1\text{res}}} \right\rfloor$
    $\textit{index} \leftarrow f_{1\text{idx}} + \textit{rangeInF1} + \left\lfloor \frac{\textit{tempCelsius} - f_{1\max}}{f_{2\text{res}}} \right\rfloor$
  else if $f_{2\max} < \textit{tempCelsius} \le f_{3\max}$ **then**
    $\textit{rangeInF1} \leftarrow \left\lfloor \frac{f_{1\max} - f_{1\min}}{f_{1\text{res}}} \right\rfloor,\quad \textit{rangeInF2} \leftarrow \left\lfloor \frac{f_{2\max} - f_{1\max}}{f_{2\text{res}}} \right\rfloor$
    $\textit{index} \leftarrow f_{1\text{idx}} + \textit{rangeInF1} + \textit{rangeInF2} + \left\lfloor \frac{\textit{tempCelsius} - f_{2\max}}{f_{3\text{res}}} \right\rfloor$
  else if $\textit{tempCelsius} > f_{3\max}$ **then**
    $\textit{rangeInF1} \leftarrow \left\lfloor \frac{f_{1\max} - f_{1\min}}{f_{1\text{res}}} \right\rfloor,\ \textit{rangeInF2} \leftarrow \left\lfloor \frac{f_{2\max} - f_{1\max}}{f_{2\text{res}}} \right\rfloor,\ \textit{rangeInF3} \leftarrow \left\lfloor \frac{f_{3\max} - f_{2\max}}{f_{3\text{res}}} \right\rfloor$
    $\textit{index} \leftarrow f_{1\text{idx}} + \textit{rangeInF1} + \textit{rangeInF2} + \textit{rangeInF3} + \left\lfloor \frac{\textit{tempCelsius} - f_{3\max}}{f_{\text{minres}}} \right\rfloor$
  else if $\textit{tempCelsius} \ge f_{0\max} \wedge \textit{tempCelsius} < f_{1\min}$ **then**
    $\textit{index} \leftarrow f_{1\text{idx}} - \left\lfloor \frac{f_{1\min} - \textit{tempCelsius}}{f_{0\text{res}}} \right\rfloor$
  else $\textit{tempCelsius} < f_{0\max}$
    $\textit{stepBelowF1min} \leftarrow \left\lfloor \frac{f_{1\min} - f_{0\max}}{f_{0\text{res}}} \right\rfloor,\ \textit{index} \leftarrow f_{1\text{idx}} - \textit{stepBelowF1min} - \left\lfloor \frac{f_{0\max} - \textit{tempCelsius}}{f_{\text{minres}}} \right\rfloor$
  end if
  return $\begin{cases} [0,0,0] & \text{if } \textit{index} < 0, \\ [255,255,255] & \text{if } \textit{index} \ge \left|\textit{Gradient}\right|, \\ \textit{Gradient}[\textit{index}] & \text{otherwise.} \end{cases}$
end function

[1] These parameters define the ranges and resolutions for the temperature-to-index mapping and can be set to match specific requirements. $f_{1\text{idx}}$ anchors $f_{1\min}$ to a particular index position.

3.C ADMRE Implementation

3.C.1 Processing Steps

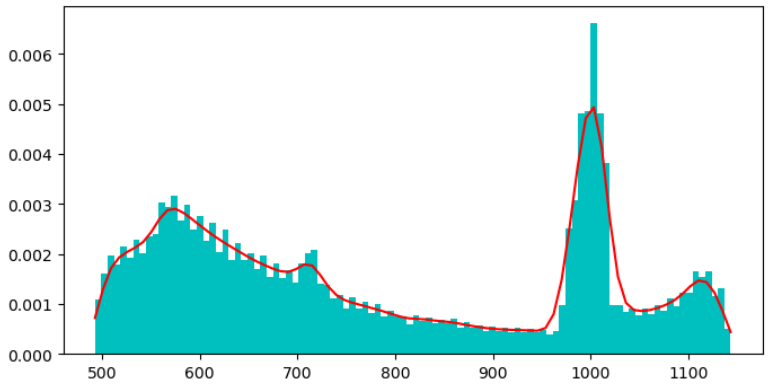


Figure 3.15: Raw data histogram, X axis with depth in mm and Y axis the pixel distribution with the KDE peak detection plotted in red.

Figure 3.16: RGB image of partially rotten green grapes.

Each detected peak is characterised by a width parameter, indicating the extent of significant depth variation, as demonstrated by the KDE-based peak detection in Figure3.15. Regions identified within these peak ranges are compressed using a finer resolution, whereas out-of-focus (OOF) areas, defined as regions closer or further than a specified focal window, are quantised more coarsely. Similarly, intervals between the identified peaks and the OOF regions ("gaps") undergo compression at a reduced resolution. This hierarchical approach prioritises depth detail around the most salient portions of the scene. Table3.8 details the region detection outcomes corresponding to Figure3.15, applying compression factors of 10 for outer regions (depth values below 500 and above 900), 3 for the gap regions, and 1 for the peak regions. The object of interest, in this case, partially rotten green grapes visible in the corresponding RGB image (Figure3.16), is centred around peak_712. By merging these individually compressed segments into a unified depth map and subsequently normalising the result, the algorithm effectively preserves critical spatial detail without incurring the computational overhead associated with uniformly high-resolution representations [55]. For visual comparison between standard normalisation and the ADMRE-processed result, Figure 3.13 shows a close-up of the grapes.

Table 3.8: Segmentation showing start/end indices and the number of unique values.

Region	Start	End	Unique Values	Compressed
oof_near	493	499	7	1
gap_1	500	509	10	4
peak_570	510	630	121	121
gap_2	631	691	61	21
peak_712	692	732	41	41
gap_3	733	803	71	24
peak_824	804	844	41	41
peak_854	845	874	30	30
gap_5	875	899	25	9
oof_far	900	1143	244	25

3.C.2 Detailed Evaluation Results

Table 3.9: Per-class mean IoU (%) for each depth encoding method combined with RGB3, along with summary metrics and estimated processing time per image (ms) for 640×480 resolution. DF980N achieves both superior segmentation accuracy and significantly lower computational cost compared to HHA.

Class	**D** (%)	**DF** (%)	**HHA** (%)	**DF980N** (%)
Lemon	59.10	60.70	56.56	**61.45**
Lemon Bad	39.54	41.22	**49.34**	40.11
Lemon Fake	6.37	9.47	10.82	**13.02**
Mirror	98.56	98.72	98.77	**98.97**
Bowl	91.24	91.72	91.20	**91.53**
Mandarin	76.83	71.70	**81.81**	78.81
Mandarin Bad	43.83	36.72	**66.29**	49.54
Mandarin Fake	**82.29**	64.80	79.40	80.53
Kettle	93.27	92.09	**94.07**	94.03
Lemon Half	41.60	38.30	35.02	**48.75**
Mandarin Half	53.84	61.31	**72.87**	70.11
Mandarin Peel	47.26	50.47	38.60	**64.40**
Cup Hot	23.31	40.62	38.63	**45.69**
Onion Red	92.77	86.24	**96.10**	93.49
Onion	96.92	96.71	97.10	**97.12**
Grapes Green	83.85	**90.28**	89.12	82.83
Grapes Green Bad	62.43	**85.84**	82.52	62.59
Grapes Green Fake	89.08	89.47	60.36	**90.64**
Grapes Blue Fake	90.56	91.72	57.57	**93.79**
Grapes Blue	61.72	92.34	86.90	**95.15**
Grapes Blue Bad	68.66	92.83	90.88	**94.94**
Apple	78.13	70.80	**84.40**	72.57
Apple Fake	**78.36**	77.53	75.58	76.40
Apple Green	79.23	81.08	74.82	**88.79**
Apple Green Bad	**75.18**	57.77	66.17	74.86
Apple Green Fake	75.44	80.58	71.34	**92.65**
Cup Cold	42.72	**46.17**	43.48	43.71
Pear	78.70	68.66	74.11	**90.08**
Pear Bad	78.01	68.67	81.47	**83.54**
Carrot	90.49	89.43	89.79	**90.70**
Carrot Fake	81.90	79.08	69.62	**82.01**
Background	99.85	99.86	99.85	**99.86**
Mean IoU	70.66	71.97	72.02	**76.33**
Freq IoU	99.14	99.17	99.16	**99.27**
Mean Pixel Acc	81.32	80.53	81.59	**84.37**
Pixel Acc	99.48	99.51	99.50	**99.57**
Mean Rank	2.95	2.72	2.70	**1.62**
processing time (ms)	0	23	250-500	25

3.C.3 ADMRE Algorithm

Algorithm 3 ADMRE: Adaptive Depth Multi-Resolution Encoding

Inputs:
raw_depth – 16-bit depth image
min_focus, **max_focus** – Minimum and maximum focus boundaries
min_width, **max_width** – Minimum and maximum peak widths for density peaks
res_oof_near, **res_oof_far**, **res_gap**, **res_focus** – Compression resolutions
num_peaks – Maximum number of dominant peaks to consider
num_channels – Output format: 1-channel (grayscale) or 2-channel (24-bit)
Output:
compressed_depth – Processed depth image with one of the following formats:
8-bit grayscale (single-channel)
24-bit (two-channel) encoded representation
function DepthFocus(raw_depth)
  Clean depth image:
    Remove NaN and zero values.
    (Optional) Remove outliers if *ol_threshold* > 0. ▷ Remove depth values with an occurrence count not higher than threshold
  Compute KDE [75, 76] over non-zero depth values:
    Adjust kernel bandwidth, e.g. `bw_method = 0.2`.
    Find peaks $(p_1, p_2, \ldots, p_{num\_peaks})$ in the probability density function, along with their widths.
  Classify regions in depth map:
    OOF near: depth $<$ *min_focus*.
    OOF far: depth $>$ *max_focus*.
    Peak regions: for each peak p_i, retain depth within $p_i \pm width_i$.
    Gaps: intervals between OOF or peak regions.
  Compress depth:
    OOF near $\rightarrow$ resolution = *res_oof_near*.
    OOF far $\rightarrow$ resolution = *res_oof_far*.
    Gaps $\rightarrow$ resolution = *res_gap*.
    Peaks $\rightarrow$ resolution = *res_focus*.
    Apply CompressDepthWithResolution(. . .) to each region accordingly.
  Merge regions into final depth map:
    Combine the compressed values for OOF, peak, and gap segments.
  Normalise and convert to desired output:
  if *num_channels* = 1 **then**:
    Scale to 0–255 for an 8-bit grayscale image.
  else if *num_channels* = 2 **then** :
    Quantise to 0–979 range.
    Map each quantised value to colour lookup (e.g., red, green channels).
    Preserve blue channel for additional features (e.g., surface normals).
  end if
  Optionally compute surface normals:
    Derive normals from final depth map.
    Encode normals as pixel values (in blue channel or separate buffer).
  return *compressed_depth*
end function

3.D MAR Implementation

3.D.1 Inverse Mapping

In the forward mapping, for each pixel at coordinates (x, y) in the source image of dimensions $w \times h$, the mapping yields target coordinates $(m_x(x, y), m_y(x, y))$, where:

$$x_n = \frac{x - c_x}{f_x}, \quad y_n = \frac{y - c_y}{f_y}, \tag{3.2}$$

and (c_x, c_y), f_x, and f_y are elements of the camera's intrinsic parameters. Although this normalisation is part of the re-distortion process, we consider only the mapping arrays, `map_x` and `map_y`, indicating the new pixel positions after distortion correction. The goal is to obtain an inverse mapping (i_x, i_y) such that, for a target pixel (u, v), if there exists a source pixel (x, y) with:

$$m_x(x, y) \approx u \quad \text{and} \quad m_y(x, y) \approx v, \tag{3.3}$$

then:

$$i_x(u, v) = x \quad \text{and} \quad i_y(u, v) = y. \tag{3.4}$$

The algorithm iterates over each pixel (x, y) in the source image, and for each, it assigns:

$$i_x(m_x(x, y), m_y(x, y)) = x, \quad i_y(m_x(x, y), m_y(x, y)) = y, \tag{3.5}$$

provided that the target coordinates $(m_x(x, y), m_y(x, y))$ fall within the bounds of the image. When multiple source pixels map to the same target coordinate, the first encountered mapping is retained. Since the forward mapping is not necessarily bijective, some target pixels may not receive any assignment due to duplication or omission in the forward process. The inverse map initially contains gaps (denoted by a placeholder value, e.g., -1). These gaps are subsequently filled using a nearest neighbour expansion method to ensure a complete inverse map, which is essential for discrete label maps and reduces the need for manual correction. This method provides a practical, approximate inversion of the forward mapping, allowing us to recover source image coordinates from the target image.

3.D.2 Re-distortion

The re-distortion is implemented as follows: initially, pixel coordinates (x, y) are converted into normalised image coordinates (x_n, y_n) using the camera's intrinsic parameters. This is identical to the inverse mapping, since the mapping process also includes a distortion correction. Therefore, Equation 3.2 is applied, where f_x and f_y are the focal lengths, and (c_x, c_y) is the principal point. The radial distance is then computed by

$$r^2 = x_n^2 + y_n^2. \tag{3.6}$$

A customised distortion model is applied to the normalised coordinates with separate asymmetry adjustments for each axis:

$$x_d = x_n \left(1 + \alpha_x \, k_1 \, r^3 + \alpha_x \, k_2 \, r^4\right), \tag{3.7}$$

$$y_d = y_n \left(1 + \alpha_y \, k_1 \, r^3 + \alpha_y \, k_2 \, r^4\right), \tag{3.8}$$

where α_x and α_y (`x_factor` and `y_factor`) are manually determined scale factors for axis-specific distortion limits, empirically set through experimentation, and k_1 and k_2 are the distortion coefficients derived from camera calibration. The distorted normalised coordinates are then converted back to pixel coordinates:

$$x_{\text{final}} = x_d \, f_x + c_x, \quad y_{\text{final}} = y_d \, f_y + c_y. \tag{3.9}$$

This mapping is applied across the image to produce the re-distorted image using the interpolation function `cv2.remap`. This formulation, which employs higher-order terms (r^3 and r^4) alongside the adjustable factors α_x and α_y, offers increased flexibility for modelling complex, non-linear distortion, particularly when the distortion is not purely radial. As a result, it effectively reverses any prior distortion introduced by alignment or calibration steps, restoring the original spatial distribution of pixels.

3.D.3 Depth Correction

Depth information is incorporated to refine the remapping. For each mapped object, the average depth is computed from the corresponding aligned depth map. This average depth value is then used to calculate correction factors that account for differences in the field of view (FOV) between the sensors, placing the labels more accurately in the target modalities. Figure 3.8a and Figure 3.8b provide a visual comparison of the annotations before and after applying the depth-based correction, demonstrating the enhanced alignment accuracy achieved by this approach.

3.D.4 Edge-Guided Random Walker Refinement

Finally, we optimise the labels by applying a random walker segmentation by first detecting edges using the Canny algorithm [64]. The detected edges are added to the seed area, and the initial RGB annotation is dilated by a factor (derived from the annotation map size) to define a maximum growth area. Subsequently, the annotation is eroded to create a consistent seed map. A Random Walker [63, 65] algorithm expands each seed region based on pixel characteristics specific to the target modality, more accurately delineating object boundaries. Figure 3.8c illustrates these seed and growth regions while Figure 3.8d shows the result. For scenes containing multiple objects, we first identify all object seed regions, excluding overlapping or occluded areas from one another's growth domain to avoid growing into other objects in low contrast scenarios. Additionally, we compute the average pixel intensity within both seed and growth areas to dynamically adjust random walker parameters, improving results across varied contrast conditions. This method performs well in scenes with sufficient contrast in the thermal or UV data, though its effectiveness may be limited in very low contrast conditions. Despite the geometric transformations and optional depth-based adjustments described, exact label alignment remains challenging when objects exhibit overlapping, complex geometries or when cameras capture significantly different viewing angles. Consequently, refinement methods such as random-walker-driven segmentation or advanced machine-learning techniques like the Segment Anything Model (SAM) [79] become essential for generating sufficiently accurate automated annotations. In our workflow, we integrated the SAM refinements within Label Studio using its SDK [80] and python, enabling us to automatically compute bounding boxes for each object and invoke SAM on a backend server to dynamically produce auto-annotations based on size, position, and class name. Although this approach can be adopted without Label Studio, the platform's flexibility and support for external models facilitated efficient iterative improvements to our automated labelling pipeline. In scenarios where SAM or similar segmentation models are employed, the random walker step may be omitted.

Table 3.10: Comparison of generated MAR labels against manually corrected thermal and UV labels (Mean IoU and Pixel Accuracy, %).

	Thermal		**UV**	
Class	IoU (%)	PixelAcc (%)	IoU (%)	PixelAcc (%)
Apple	89.94	93.31	87.31	90.67
Apple Fake	87.65	91.10	87.26	90.72
Apple Green	87.59	91.44	83.34	88.04
Apple Green Bad	86.39	87.80	83.59	88.93
Apple Green Fake	88.21	91.80	87.24	91.70
Bowl	85.51	88.60	77.23	83.07
Carrot	84.21	92.45	76.11	87.43
Carrot Fake	63.80	83.26	60.12	78.39
Cup Cold	89.78	81.45	73.10	78.59
Cup Hot	84.37	90.83	84.52	91.00
Grapes Blue	87.47	88.52	86.42	90.32
Grapes Blue Bad	87.95	89.64	90.05	91.81
Grapes Blue Fake	87.08	87.65	83.51	87.31
Grapes Green	88.25	91.48	81.18	87.68
Grapes Green Bad	85.37	86.72	81.02	86.09
Grapes Green Fake	86.66	89.67	79.88	86.24
Kettle	80.35	86.13	74.19	83.56
Lemon	85.27	90.70	81.32	87.30
Lemon Bad	86.55	91.44	80.14	84.48
Lemon Fake	84.51	91.33	79.02	86.20
Lemon Half	86.14	91.91	84.73	88.75
Mandarin	85.43	92.47	83.23	89.53
Mandarin Bad	83.11	92.30	76.95	86.13
Mandarin Fake	87.78	91.82	87.06	92.26
Mandarin Half	82.98	88.94	85.61	90.00
Mandarin Peel	79.88	90.98	78.56	83.59
Mirror	87.18	98.18	86.50	94.54
Onion	87.67	90.93	88.82	93.43
Onion Red	86.69	91.00	86.78	92.44
Pear	76.63	86.19	82.04	88.18
Pear Bad	78.54	91.06	83.23	88.91
ALL	84.81	90.04	81.94	87.98

3.D.5 Flowchart

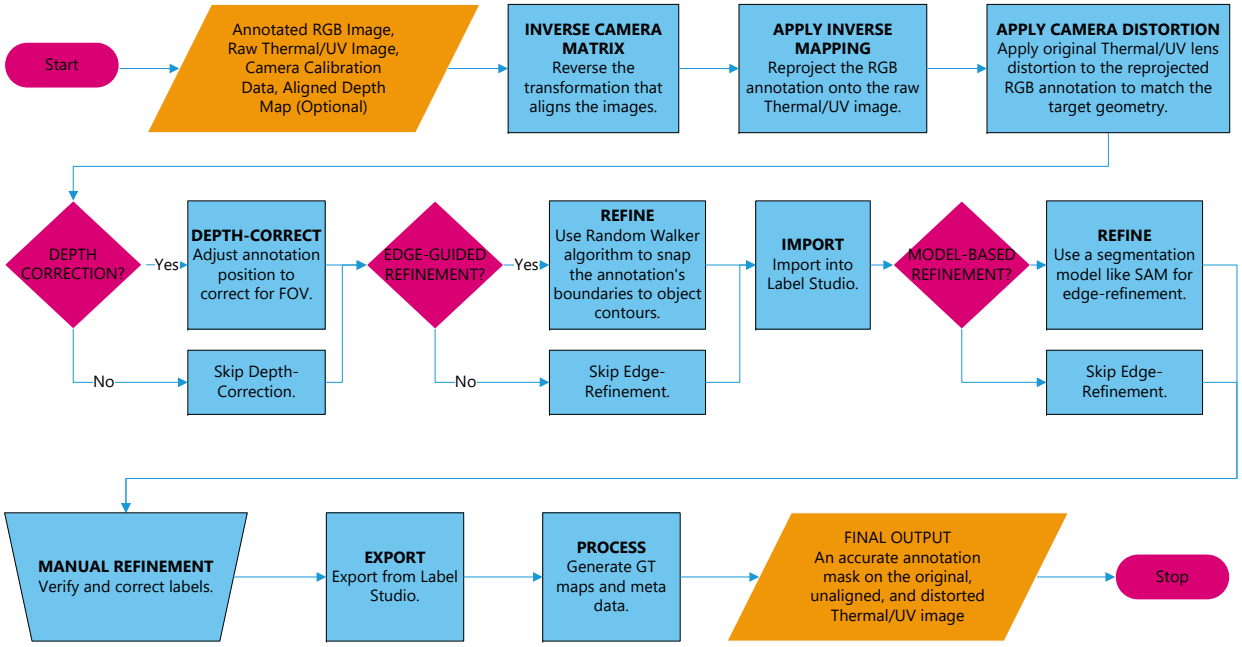


Figure 3.17: MAR Flowchart

3.D.6 MAR Algorithm

Algorithm 4 MAR: Multimodal Annotation Remapping

Inputs:
rgb_image, rgb_anno – Original RGB image and corresponding label or instance mask
thermal_image_raw, uv_image_raw – Unprocessed thermal and UV images
camera_intrinsics (fx, fy, cx, cy) and **extrinsics** (R, T) for each modality
depth_map – (optional) Depth data aligned to the RGB image
canny_thresholds, random_walker_params – Parameters for edge detection and region-growing refinement
Output:
thermal_anno, uv_anno – Final labels on thermal and UV images
function ReverseMapAnnotations(rgb_anno, camera_intrinsics, extrinsics, depth_map) ▷ MAR pipeline
  Derive or load `map_x`, `map_y` for forward mapping from RGB to target modality (e.g. thermal)
  Initialize `inv_map_x`, `inv_map_y` with -1 ▷ Placeholder values for inverse map
  for $x \in [0, \textit{width\_rgb}]$, $y \in [0, \textit{height\_rgb}]$ **do**
    $(t\_x, t\_y) \leftarrow \big(\texttt{map\_x}[x, y],\ \texttt{map\_y}[x, y]\big)$
    if (t_x, t_y) within target (thermal/UV) image bounds **then**
      if $\texttt{inv\_map\_x}[t\_x, t\_y] == -1$ **then**
        $\texttt{inv\_map\_x}[t\_x, t\_y] \leftarrow x; \quad \texttt{inv\_map\_y}[t\_x, t\_y] \leftarrow y$
      end if
    end if
  end for
  FillGaps(`inv_map_x`, `inv_map_y`) ▷ Nearest-neighbor or similar to fill holes in inverse map
  x_anno_remapped ← Remap(`inv_map_x`, `inv_map_y`, *rgb_anno*) ▷ Applies inverse map
  x_anno_distorted ← ApplyReDistortion(*x_anno_remapped*, *camera_intrinsics*, *distortionParams*)
  if *depth_map* **is available then**
    x_anno_corrected ← DepthCorrection(*x_anno_distorted*, *depth_map*)
  else
    x_anno_corrected ← *x_anno_distorted*
  end if
  x_edges ← CannyEdgeDetection(*x_image_raw*, *canny_thresholds*)
  x_anno_final ← RefineWithRandomWalker(*x_anno_corrected*, *x_edges*, *random_walker_params*)
  Return *x_anno_final*
end function
function DepthCorrection(annotation, depth_map) ▷ Optional FOV scaling using average instance depth
  for *instance* ∈ *annotation* **do**
    avgDepth ← ComputeAverageDepth(*instance*, *depth_map*)
    scaleFactor ← CalcFOVAdjustment(*avgDepth*, *cameraParams*)
    instance ← ScaleAnnotation(*instance*, *scaleFactor*)
  end for
  Return *annotation*
end function
function RefineWithRandomWalker(annotation, edges, params)
  seedMask ← Dilate($annotation$, *params.dilationFactor*)
  seedMask ← Erode(*seedMask*, *params.erosionFactor*)
  seedMask ← *seedMask* + *edges*
  refinedMask ← RandomWalkerSegmentation(*seedMask*, *edges*, *params*)
  Return *refinedMask*
end function

Bibliography

[1] M. Brenner, N. H. Reyes, T. Susnjak, and A. L. C. Barczak. RGB-D and thermal sensor fusion: A systematic literature review. *IEEE Access*, 11:102667–102685, 2023.

[2] Jon Muhovič and Janez Perš. Joint calibration of a multimodal sensor system for autonomous vehicles. *Sensors*, 23(12), 2023.

[3] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

[4] M. Abdul-Al, G. Kumi Kyeremeh, R. Qahwaji, N. T. Ali, and R. A. Abd-Alhameed. The Evolution of Biometric Authentication: A Deep Dive Into Multi-Modal Facial Recognition: A Review Case Study. *IEEE Access*, 12:179010–179038, 2024.

[5] Kechen Song, Jie Wang, Yanqi Bao, Liming Huang, and Yunhui Yan. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. *IEEE/ASME Transactions on Mechatronics*, 28(3):1558–1569, 2022.

[6] Hongwei Wen, Kechen Song, Liming Huang, Han Wang, Junyi Wang, and Yunhui Yan. Hierarchical two-stage modal fusion for triple-modality salient object detection. *Measurement*, 218:113180, 2023.

[7] Kechen Song, Han Wang, Ying Zhao, Liming Huang, Hongwen Dong, and Yunhui Yan. Lightweight multi-level feature difference fusion network for RGB-D-T salient object detection. *Journal of King Saud University - Computer and Information Sciences*, 35(10):101702, 2023.

[8] Bin Wan, Xiaofei Zhou, Yaoqi Sun, Zunjie Zhu, Hongkui Wang, Chenggang Yan, et al. Tmnet: Triple-modal interaction encoder and multi-scale fusion decoder network for vdt salient object detection. *Pattern Recognition*, 147:110074, 2024.

[9] Liuxin Bao, Xiaofei Zhou, Xiankai Lu, Yaoqi Sun, Haibing Yin, Zhenghui Hu, Jiyong Zhang, and Chenggang Yan. Quality-aware selective fusion network for vdt salient object detection. *IEEE Transactions on Image Processing*, 33:3212–3226, 2024.

[10] Jiyuan Qiu, Chen Jiang, and Haowen Wang. ETFormer: An Efficient Transformer Based on Multimodal Hybrid Fusion and Representation Learning for RGB-D-T Salient Object Detection. *IEEE Signal Processing Letters*, 31:2928–2932, 2024.

[11] Nianchang Huang, Yang Yang, Ruida Xi, Qiang Zhang, Jungong Han, and Jin Huang. Salient Object Detection From Arbitrary Modalities. *arXiv preprint arXiv:2405.03352*, 2024. Under review.

[12] Martin Brenner, Napoleon Reyes, Teo Susnjak, and Andre Barczak. MM5: Multimodal Image Dataset, 2025. Dataset.

[13] C. Palmero, A. Clapés, C. H. Bahnsen, A. Møgelmose, T. B. Moeslund, and S. Escalera. Multi-modal RGB–Depth–Thermal human body segmentation. *Int. J. Computer Vision*, 118(2):217–239, 2016.

[14] Christian Stippel, Thomas Heitzinger, and Martin Kampel. A trimodal dataset: Rgb, thermal, and depth for human segmentation and temporal action detection. In *DAGM German Conference on Pattern Recognition*, pages 18–33. Springer, 2023.

[15] Vladimir V. Kniaz, Vladimir A. Knyaz, Jiří Hladůvka, Walter G. Kropatsch, and Vladimir Mizginov. ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Computer Vision – ECCV 2018 Workshops*, volume 11134 of *Lecture Notes in Computer Science*, pages 606–624. Springer, Cham, 2019.

[16] Yuanyuan Shi, Yunan Li, Siyu Liang, Huizhou Chen, and Qiguang Miao. MGR-Dark: A large multimodal video dataset and RGB-IR benchmark for gesture recognition in darkness. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 2321–2330. Association for Computing Machinery, 2024.

[17] Dayan Guan, Yanpeng Cao, Jun Liang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019.

[18] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

[19] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022.

[20] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.

[21] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, 99:101870, 2023.

[22] Albert Mosella-Montoro and Javier Ruiz-Hidalgo. 2d–3d geometric fusion network using multi-neighbourhood graph convolution for rgb-d indoor scene classification. *Information Fusion*, 76:46–54, 2021.

[23] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012.

[24] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proc. CVPR*, pages 1037–1045, 2015.

[25] Z. Zhang, J. Yan, and S. Liu. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proc. CVPR*, pages 919–928, 2019.

[26] Gianni Franchi, Marwane Hariat, Xuanlong Yu, Nacim Belkhir, Antoine Manzanera, and David Filliat. InfraParis: A multi-modal and multi-task autonomous driving dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2973–2983. IEEE, 2024.

[27] Michael Baltaxe, Tomer Pe'er, and Dan Levi. Polarimetric imaging for perception. In *Proceedings of the 34th British Machine Vision Conference (BMVC)*, page 566. British Machine Vision Association, 2023.

[28] J. Yin, A. Li, T. Li, W. Yu, and D. Zou. M2DGR: A multi-sensor and multi-scenario SLAM dataset for ground robots. *IEEE Robotics Autom. Lett.*, 7(2):2266–2273, 2022.

[29] Hanzhe Teng, Yipeng Wang, Xiaoao Song, and Konstantinos Karydis. Multimodal dataset for localization, mapping and crop monitoring in citrus tree farms. In George Bebis, Golnaz Ghiasi, Yi Fang, Andrei Sharf, Yue Dong, Chris Weaver, Zhicheng Leo, Joseph J. LaViola Jr., and Luv Kohli, editors, *Advances in Visual Computing*, pages 571–582, Cham, 2023. Springer Nature Switzerland.

[30] Mikkel Fly Kragh, Peter Christiansen, Morten Stigaard Laursen, Morten Larsen, Kim Arild Steen, Ole Green, Henrik Karstoft, and Rasmus Nyholm Jørgensen. Fieldsafe: Dataset for obstacle detection in agriculture. *Sensors*, 17(11), 2017.

[31] Mahmoud Abdulsalam, Zakaria Chekakta, Nabil Aouf, and Maxwell Hogan. Fruity: a multi-modal dataset for fruit recognition and 6d-pose estimation in precision agriculture. In *2023 31st Mediterranean Conference on Control and Automation (MED)*, pages 144–149. IEEE, 2023.

[32] Pedro J Navarro, Leanne Miller, María Victoria Díaz-Galián, Alberto Gila-Navarro, Diego J Aguila, and Marcos Egea-Cortines. A novel ground truth multispectral image dataset with weight, anthocyanins, and brix index measures of grape berries tested for its utility in machine learning pipelines. *GigaScience*, 11:giac052, 2022.

[33] Kechen Song, Jie Wang, Yanqi Bao, Liming Huang, and Yunhui Yan. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. *IEEE/ASME Transactions on Mechatronics*, 2022.

[34] Julian Strohmayer and Martin Kampel. A compact tri-modal camera unit for rgbdt vision. In *2022 the 5th International Conference on Machine Vision and Applications (ICMVA)*, ICMVA 2022, page 34â€“42, 2022.

[35] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[36] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, AndrÃ© Anjos, and SÃ©bastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2019.

[37] Anjith George and SÃ©bastien Marcel. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 16:361–375, 2020.

[38] Scott M. Pizer, E. P. Amburn, J. D. Austin, Robert Cromartie, Allen Geselowitz, Terence Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Comput. Vision Graph. Image Process.*, 39(3):355–368, 1987.

[39] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In Paul S. Heckbert, editor, *Graphics Gems IV*, pages 474–485. Academic Press, San Diego, CA, 1994.

[40] Huda I. Ashiba, Hala M. Mansour, Hossameldin M. Ahmed, Mahmoud I. Dessouky, Mahmoud F. El-Kordy, Osama Zahran, and Fathi E. Abd El-Samie. Enhancement of infrared images using histogram processing and the undecimated additive wavelet transform. *Multimedia Tools Appl.*, 78(9):11277–11290, 2019.

[41] Volker Schatz. Low-latency histogram equalization for infrared image sequences: a hardware implementation. *J. Real-Time Image Process.*, 8:193–206, 2013.

[42] V. E. Vickers. Plateau equalization algorithm for real-time display of high-quality infrared imagery. *Opt. Eng.*, 35(7):1921–1926, 1996.

[43] Manash P. Das, Larry H. Matthies, and Shreyansh Daftry. Online photometric calibration of automatic gain thermal infrared cameras. *IEEE Robot. Autom. Lett.*, 6(2):2453–2460, 2021.

[44] Wojciech Jamrozik, Jacek G
'orka, and Gilmar F. Batalha. Dynamic range compression of thermograms for assessment of welded joint face quality. *Sensors*, 23(4):1995, 2023.

[45] Dong-Guw Lee, Jeongyun Kim, Younggun Cho, and Ayoung Kim. Thermal chameleon: Task-adaptive tone-mapping for radiometric thermal-infrared images. *arXiv:2410.18340 [cs.CV]*, 2024.

[46] Axel Gödrich, Daniel König, Gabriel Eilertsen, and Michael Teutsch. Joint tone mapping and denoising of thermal infrared images via multi-scale retinex and multi-task learning. In *Infrared Technology and Applications XLIX (Proc. SPIE)*, volume 12534, page 1253417, 2023.

[47] Huaizhou Li, Shuaijun Wang, Sen Li, Hong Wang, Shupei Wen, and Fengyu Li. Thermal infrared image enhancement algorithm based on multi-scale guided filtering. *Fire*, 7(6):192, 2024.

[48] Youpan Zhu, Yongkang Zhou, Wenyao Jin, Lujun Zhang, Guisheng Wu, and Yufang Shao. A low-delay dynamic range compression and contrast enhancement algorithm based on an uncooled infrared sensor with local optimal contrast. *Sensors*, 23(21):8860, 2023.

[49] Marc Miró Duch, Josep R. Morros, and Javier Ruiz-Hidalgo. Depth map compression via 3d region-based representation. *Multimedia Tools and Applications*, 76(11):13761–13784, 2017.

[50] T. Sonoda and A. Grunnet-Jepsen. Depth image compression by colorization for intel realsense depth cameras. `https://dev.intelrealsense.com/docs/depth-image-compression-by-colorization-for-intel-realsense-depth-cameras`, 2021. Intel RealSense White Paper.

[51] Jason Rambach, Bruno Mirbach, Yuriy Anisimov, and Didier Stricker. Time-of-flight depth sensing for automotive safety and smart building applications: The VIZTA project. *IEEE Access*, 11:105819–105829, 2023.

[52] Miaoqing Chen, Peng Zhang, Zhan Chen, Yebin Zhang, Xun Wang, and Sam Kwong. End-to-end depth map compression framework via rgb-to-depth structure priors learning. In *IEEE International Conference on Image Processing (ICIP)*, pages 3206–3210, Bordeaux, France, 2022.

[53] Jean-Philippe D'Amato. FitDepth: Fast and lite 16-bit depth image compression algorithm. *EURASIP J. Image Video Process.*, 2023:5, 2023.

[54] Andrew D. Wilson. Fast lossless depth image compression. In *Proc. ACM International Conference on Interactive Surfaces and Spaces (ISS)*, pages 100–105, 2017.

[55] Mohammad Ali Tahouri, Alin Adrian Alecu, Leon Denis, and Adrian Munteanu. Lossless and near-lossless l_∞ compression of depth video data. *Sensors*, 25(5):1403, 2025.

[56] Radhakrishnan Gopalapillai, Deepa Gupta, Mohammed Zakariah, and Yousef A. Alotaibi. Convolution-based encoding of depth images for transfer learning in rgb-d scene classification. *Sensors*, 21(23):7950, 2021.

[57] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision (ECCV)*, pages 345–360. Springer, 2014.

[58] S. Hareesh Kumar and K. R. Ramakrishnan. Depth compression via planar segmentation. *Multimedia Tools and Applications*, 78(5):6529–6558, 2019.

[59] Fang Tan, Zhaoqiang Xia, Yupeng Ma, and Xiaoyi Feng. 3d sensor based pedestrian detection by integrating improved HHA encoding and two-branch feature fusion. *Remote Sensing*, 14(3):645, 2022.

[60] Pietro Ruiu, Lorenzo Mascia, and Enrico Grosso. Saliency-guided point cloud compression for 3d live reconstruction. *Multimodal Technologies and Interaction*, 8(5):36, 2024.

[61] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[62] G. Bradski and A. Kaehler. OpenCV library. `https://opencv.org/`, 2000. Accessed: March 22, 2023.

[63] Leo Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

[64] John F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[65] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: Image processing in python, 2014. Version 0.10.0.

[66] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[67] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[68] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[69] Maria João Sousa, Alexandra Moutinho, and Miguel Almeida. Thermal infrared sensing for near real-time data-driven fire detection and monitoring systems. *Sensors*, 20(23):6803, 2020.

[70] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023.

[71] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

[72] Virgil E. Vickers. Plateau equalization algorithm for real-time display of high-quality infrared imagery. *Optical Engineering*, 35(7):1921–1926, 1996.

[73] Daniel J. Jobson, Zia-ur Rahman, and Glenn A. Woodell. A multiscale retinex for color image enhancement. *IEEE transactions on image processing*, 6(7):965–976, 1997.

[74] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[75] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[76] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[77] Fang Tan, Zhaoqiang Xia, Yupeng Ma, and Xiaoyi Feng. 3d sensor based pedestrian detection by integrating improved hha encoding and two-branch feature fusion. *Remote Sensing*, 14(3), 2022.

[78] Label Studio Team. Label Studio: Open Source Data Labelling Platform, 2024. Accessed: 8 Feb. 2025.

[79] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Trevor Xiao, Spencer Whitehead, Alexander Berg, Wan-Yen Lo, Piotr Dollár, and Kaiming He. Segment anything. *arXiv preprint*, arXiv:2304.02643, 2023.

[80] HumanSignal. Label studio sdk: Python client for label studio api. `https://github.com/HumanSignal/label-studio-sdk`, 2025. Accessed: April 4, 2025.

Chapter 4

Encoder-Level Fusion for Semantic Segmentation

With the MM5 dataset and preprocessing tools established in Chapter 3, this chapter addresses the architectural challenge of efficiently fusing all five modalities. The systematic review (Chapter 2) highlighted the absence of real-time transformer-based architectures for tri-modal or higher fusion, while MM5 provides the first benchmark enabling such experimentation with RGB, depth, infrared intensity, thermal, and ultraviolet data. This chapter introduces GatedFusion-Net, an encoder-level fusion architecture that integrates all five aligned modalities through stage-wise geometric enhancement and per-pixel gating mechanisms. The architecture leverages MM5's preprocessed depth (ADMRE) and thermal (DTMRE) representations to achieve real-time performance while establishing reproducible baselines. The failure analysis presented in this chapter subsequently motivates the decoder-level approach developed in Chapter 5, as encoder gates are found to behave as static, lighting-specific weights that degrade substantially under unexpected modality loss.

The contents of this chapter are reproduced from the following article:

Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2026). GatedFusion-Net: Per-pixel modality weighting in a five-cue transformer for RGB-D-I-T-UV fusion. Information Fusion*, 129, 103986. https://doi.org/10.1016/j.inffus.2025.103986*.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Martin Brenner
Name and title of main supervisor:	Dr Napoleon Reyes
In which chapter is the manuscript/published work?	4

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

The candidate was the main contributor of this work, and has done the literature review, experiments, and drafted the manuscript. The final draft was completed with the suggestions from the co-authors.

Please select one of the following three options:

(●) **The manuscript/published work is published or in press**

Please provide the full reference of the research output:

Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2025). GatedFusion-Net: Per-Pixel Modality Weighting in a Five-Cue Transformer For RGB-D-I-T-UV Fusion. Information Fusion, 103986. https://doi.org/https://doi.org/10.1016/j.inffus.2025.103986

( ) **The manuscript is currently under review for publication**

Please provide the name of the journal:

( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Digitally signed by Martin Brenner DN: cn=Martin Brenner, c=NZ, email=mb@fisag.co.nz Reason: I agree to specified portions of this document Location: Auckland Date: 2025.11.25 20:47:40 +13'00'	Main supervisor's signature:	Napoleon Reyes Digitally signed by Napoleon Reyes Date: 2025.12.01 17:09:36 +13'00'

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

4.1 Abstract

We introduce GatedFusion-Net (GF-Net), built on the SegFormer Transformer backbone, as the first architecture to unify RGB, depth (D), infrared intensity (I), thermal (T), and ultraviolet (UV) imagery for dense semantic segmentation on the MM5 dataset. GF-Net departs from the CMX baseline via: (1) stage-wise RGB–intensity–depth enhancement that injects geometrically aligned D, I cues at each encoder stage, together with surface normals (N), improving illumination invariance without adding parameters; (2) per-pixel sigmoid gating, where independent Sigmoid Gate blocks learn spatial confidence masks for T and UV and add their contributions to the RGB+DIN base, trimming computational cost while preserving accuracy; and (3) modality-wise normalisation using per-stream statistics computed on MM5 to stabilise training and balance cross-cue influence. An ablation study reveals that the five-modality configuration (RGB+DIN+T+UV) achieves a peak mean IoU of 88.3%, with the UV channel contributing a 1.7-percentage-point gain under optimal lighting (RGB3). Under challenging illumination, it maintains comparable performance, indicating complementary but situational value. Modality-ablation experiments reveal strong sensitivity: removing RGB, T, DIN, or UV yields relative mean IoU reductions of 83.4%, 63.3%, 56.5%, and 30.1%, respectively. Sigmoid-Gate fusion behaves primarily as static, lighting-dependent weighting rather than adapting to sensor loss. Throughput on an RTX 3090 with a MiT-B0 backbone is real-time: 640×480 at 74 fps for RGB+DIN+T, 55 fps for RGB+DIN+T+UV, and 41 fps with five gated streams. These results establish the first RGB–D–I–T–UV segmentation baselines on MM5 and show that per-pixel sigmoid gating is a lightweight, effective alternative to heavier attention-based fusion.

4.2 Introduction

Robust semantic segmentation in service robotics and automated inspection demands tolerance to challenging conditions such as poor illumination, specularities, and spectral camouflage. While conventional RGB sensing falters in these scenarios, complementary sensors can mitigate specific weaknesses: depth provides geometry independent of colour, thermal imaging highlights heat-emitting regions, infrared intensity broadens the dynamic range, and ultraviolet (UV) reveals surface fluorescence [1]. For example, household robots tasked with kitchen assistance must distinguish genuine fruit and vegetables from replicas or synthetic models, and, under variable lighting, recognise rotting or spoiled produce, while agricultural inspection systems in packing lines require reliable detection of bruising, lesions or ripeness defects on farm produce. Exploiting this sensor heterogeneity thus promises finer delineation of objects and enhanced reliability across domestic, industrial and agricultural contexts. However, the effective integration of numerous data streams within a single neural network remains a significant hurdle. For instance, in the MM5 corpus, three of the five streams originate from a factory-calibrated RGB-D sensor. In contrast, the thermal and UV cameras possess different resolutions and fields of view. Early data-level fusion approaches, which concatenate raw modality channels, are often brittle and typically necessitate perfect registration of the modalities. Consequently, recent research has focused on introducing explicit alignment or attention modules to rectify cross-modal discrepancies [2, 3]. Although transformer frameworks like CMX [2] can rectify bimodal features online, their extension to four or more aligned inputs has not yet been demonstrated. Furthermore, the diagnostic value of each cue varies across an image: depth perception is unreliable for thin structures, thermal saturation can occur under direct sunlight, and the utility of 365 nm UV in produce inspection remains uncertain. Existing fusion architectures often presuppose perfectly aligned RGB-D or RGB-T inputs [4] or delegate calibration to a preprocessing stage [5]. When more modalities are involved, mid-level attention-based fusion can become computationally intensive and susceptible to low-quality features [6, 7]. To address these challenges, we propose a transformer architecture that extends CMX to accommodate four or more modalities through three key design choices. Firstly, we implement Stage-Wise Intensity Fusion (SWIF) enhancement: geometrically paired depth (D) and infrared intensity (I) are merged at the data level and enriched with surface normals. The resultant intensity map reweights RGB features before any cross-modal interaction, yielding a texture-geometry backbone that is both compact and illumination-invariant. Secondly, we adopt the learnable Feature-Rectify module from CMX to align the thermal (T) and UV streams within each encoder stage, thereby obviating the need for external warping. Thirdly, we replace Feature Fusion Modules with per-pixel sigmoid gating, allowing thermal and UV contributions to be modulated by spatially varying confidence masks rather than global channel attention, which suppresses noise where a modality is uninformative. All inputs undergo modality-wise normalisation, computed over the MM5 dataset, to equalise dynamic ranges, a feature absent in the original CMX. Ablation experiments (Table 4.6

and 4.7) confirm that the proposed network establishes the first reproducible benchmark on MM5; across diverse indoor lighting conditions the four-modality combination of RGB, depth, intensity, and thermal imagery already captures the dominant information, while the inclusion of UV yields only modest average gains yet markedly improves some of the *bad* and *fake* subclasses, sharpening the separation of rotten or replica fruit from genuine produce. In addition, we conduct an extensive failure analysis that demonstrates the architecture's robustness during standard operations with a 99.7% pixel accuracy rate. However, it also highlights weaknesses when sensors degrade. The findings reveal that the gating mechanism focuses on static lighting-dependent strategies rather than dynamic fusion, offering clear recommendations for future enhancements in architectural design.

4.2.1 Key Contributions

The primary contributions presented in this paper are:

1. The first transformer-based segmentation architecture that integrates RGB, depth, infrared intensity, thermal, and ultraviolet modalities through a novel staged fusion framework, enabling effective multi-sensor fusion within a unified, real-time capable model.

2. A dual-stage encoder incorporating (i) stage-wise RGB–intensity-depth enhancement, (ii) modality-wise normalisation for training stability and balanced feature scaling, and (iii) pixel-level sigmoid gating that performs learnt, content-conditioned weighting of auxiliary modalities without relying on heavy attention mechanisms.

3. A comprehensive evaluation establishing the inaugural RGB-D-I-T-UV baseline on the challenging MM5 dataset, demonstrating state-of-the-art accuracy and robustness across varied lighting conditions.

4. An ablation study investigating the specific contribution of ultraviolet cues, providing the first empirical evidence regarding their complementary value relative to high-quality RGB-D-I-T data under realistic and adverse lighting scenarios.

5. A systematic comparative analysis of fusion strategies, including early (data-level) fusion, feature-level fusion, stage-wise enhancement, per-pixel gating, and channel-wise cross-modal attention-based fusion (FFM), quantifying their comparative strengths and robustness for multimodal feature integration.

4.3 Related Work

Multimodal fusion of RGB, depth and thermal cues has proven effective for robust perception under adverse conditions. A recent review by Brenner *et al.* [8] highlights that combining geometric and thermal signatures with colour appearance overcomes limitations of single-sensor systems, but progress was hindered by a lack of large, aligned 3-modal datasets. The VDT-2048 corpus [9], with its spatially registered RGB-D-T frames and the HWSI fusion scheme, marked a turning point for saliency detection, though its use of 8-bit depth and auto-gain thermal images limits low-level fusion research. Early CNN-based triple-modality networks fused separate backbones via attention or concatenation. Wen *et al.* proposed a hierarchical two-stage fusion for RGB-D-T saliency, first predicting modality-specific maps, then refining across streams [10]. Song *et al.*'s MFDF-Net achieved over 120 fps with only 8.9 M parameters by asymmetrically fusing MobileNetV2 encoders and dedicated CME/CMF modules [6]. TMNet extended this by introducing dense cross-modal interaction units atop a VGG-16 backbone, reaching state-of-the-art accuracy at 5.9 fps on 353×352 inputs [11]. Bao *et al.*'s QSF-Net addressed sensor unreliability with a quality-aware gating mechanism, adaptively downweighting noisy depth or thermal regions and achieving 11.2 fps on 384×384 inputs [7]. The advent of transformers has enabled more unified fusion. Qiu *et al.*'s ETFormer replaces multiple CNN streams with a single transformer encoder pretrained on a large synthetic RGB-D-T dataset, and a multimodal multi-head attention block, delivering richer long-range interactions and about 35 fps on 224×224 inputs [3]. Huang *et al.* further generalised this concept with their Modality Switch Network, which uses learnt modality tokens and a dynamic fusion transformer to accept any combination of RGB, depth and thermal inputs within one model, achieving flexible saliency mapping at roughly 18 fps on 224×224 inputs [5]. In summary, the past three years have witnessed rapid advances in multimodal fusion, which integrates three or more sensing channels, particularly RGB, depth, and thermal infrared. Early studies validated the value of triple-modality data through new datasets and CNN-based fusion

Table 4.1: Comparison of Multimodal Fusion Methods

Method	Modalities	Stage	Mechanism	Input Size	Real-Time	Dataset
Ozcan & Cetin (2022) [12]	RGB, D, T	Early	Alignment + stacking	$640 \times 480/320 \times 240$	Yes (≈50 fps)	own
HWSI-Net (2022) [9]	RGB, D, T	Mid	Multi-attention	352×352	No (≈3.6 fps)	VDT-2048
MFDF-Net (2023) [6]	RGB, D, T	Mid	Feature-diff fusion	320×320	Yes (124 fps)	VDT-2048
TMNet (2024) [11]	RGB, D, T	Mid	Interaction units + Attention	352×352	No (≈6 fps)	VDT-2048
QSF-Net (2024) [7]	RGB, D, T	Multi	Quality-aware fusion	384×384	No (11.16 fps)	VDT-2048
ETFormer (2024) [3]	RGB, D, T	Mid	Transformer attention	224×224	Yes (≈35 fps)	VDT-2048
AM-SOD (2024) [5]	1-3 (extensible)	Mid	Dynamic fusion	224×224	No (≈18 fps for 3 mods)	AM-XD
GF-Net MiT-B0 (2025) (Ours)	RGB, D, I, T, UV	Early/Mid	Staged fusion	$640 \times 480/320 \times 240$	Yes ($55/91$ fps)	MM5

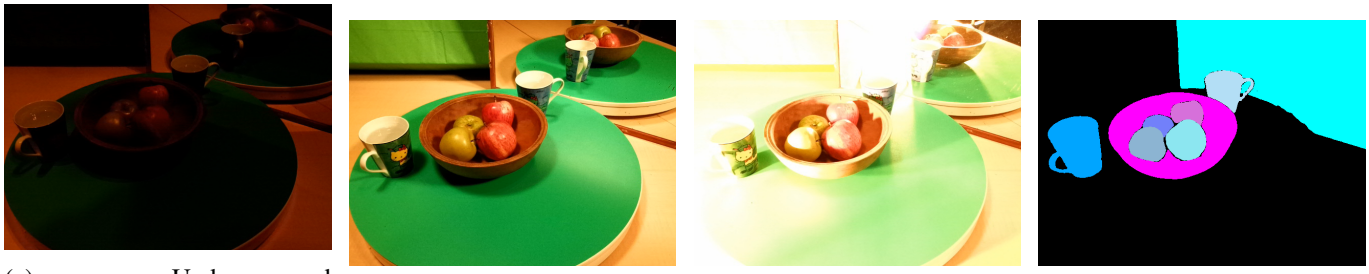
models [8, 9, 10], while more recent works utilise transformer-based encoders and quality-aware weighting to achieve state-of-the-art accuracy with improved efficiency [11, 7, 3]. Yet the community still lacks benchmarks for fusion methods beyond RGB-D or RGB-T pairs and RGB-D-T triplets. The MM5 dataset [1, 13] begins to address this gap by providing extensive raw and preprocessed imagery alongside both aligned and unaligned annotations, thereby enabling comprehensive evaluation of multimodal fusion strategies. Building on these foundations, our work proposes a novel hierarchical fusion framework that explicitly leverages the complementary strengths of five aligned modalities (RGB, depth, infrared intensity, thermal, and ultraviolet) to deliver accurate segmentation at real-time rates in complex visual scenes. Table 4.1 summarises representative multimodal fusion networks.
These recent advances, together with the emergence of MM5, motivate a set of unresolved research questions that underpin the present work:

- How can transformer-based architectures be adapted to effectively integrate more than three aligned modalities, including RGB, depth, intensity, thermal, and ultraviolet, within a unified framework?
- Can a generalisable and computationally efficient architecture be established as a strong baseline for MM5, enabling fair comparison and future development of multimodal fusion networks?
- To what degree does including ultraviolet cues improve segmentation performance under realistic, variable lighting, or do they remain redundant when high-quality RGB-D-I-T data is available?
- What is the impact of early versus feature fusion strategies, and to what extent does stage-wise enhancement of geometric cues (D+I) improve segmentation over gated or attention-based fusion?
- How does per-pixel, content-adaptive gating compare to channel-wise cross-modal fusion (FFM) for robustly integrating multimodal features, particularly in the presence of noise and redundancy?

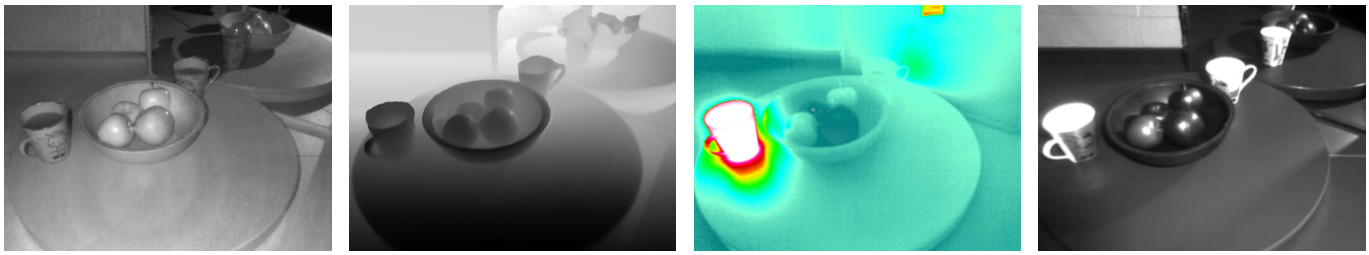
In the following sections, we detail our proposed method, which is designed to address these questions systematically.

4.4 MM5 Dataset

The MM5 dataset [1] was designed to address key limitations in existing multimodal benchmarks, which often lack sufficient modality diversity, raw sensor fidelity, and raw data annotations. MM5 systematically integrates five core imaging modalities, RGB, depth (D), thermal (T), ultraviolet (UV), and near-infrared (NIR) in a unified acquisition and annotation framework. The acquisition platform is built from off-the-shelf RGB-D components, supplemented with thermal and UV sensors. Each scene is captured under diverse lighting conditions (shadows, dim lighting, overexposure) and includes a broad range of real and replica produce, as well as partially decayed items, ensuring that each modality provides unique and sometimes complementary cues. Crucially, MM5 preserves raw 16-bit depth and thermal data, enabling advanced preprocessing and denoising studies beyond the limitations of 8-bit AGC images. The dataset provides both aligned and unaligned annotations, promoting flexibility in method development. Initial experiments using a transformer-based segmentation network on MM5 demonstrate that modality-specific preprocessing significantly improves segmentation accuracy for depth and thermal encoding.
A sample subset of images taken from the MM5 dataset, as used in later experiments, is shown in Figure 4.1. While the dataset contains eight variants of RGB images, we have focused our experiments on the underexposed, well-exposed, and overexposed images for clarity in investigating the fusion of additional modalities and their impact under these lighting conditions.



(a) Under-exposed RGB(RGB1) (b) Ideal RGB(RGB3) (c) Over-exposed RGB(RGB5) (d) Ground truth mask



(e) Infrared intensity (f) Depth (g) Thermal (T24) (h) Ultraviolet (U8)

Figure 4.1: MM5 sample subset for frame 257

4.4.1 Training and Evaluation Data

Table 4.2 summarises the class-wise MM5 train-evaluation split, as specified by the files `list_train_f.txt` and `list_eval_f.txt` included with the dataset. Except for a few sparsely represented categories, the protocol maintains a broadly stratified allocation, with 75-80% of the images reserved for training and the remainder for evaluation. This strategy maximises the amount of data available for optimisation while still providing a statistically meaningful hold-out subset for each class. The class distribution nevertheless remains skewed in absolute terms. Core fruit classes, such as Lemon and Mandarin, contribute upwards of one hundred labelled object instances (spanning 69 and 47 images, respectively), whereas others, such as Mandarin Peel and Kettle, are limited to a dozen or fewer images. Mandarin Peel remains minimally represented, and its evaluation set comprises only three images. For some underrepresented categories, the nominal 20-25% evaluation split is maintained, yet for others (e.g., Mirror), the ratio is more markedly imbalanced due to the dataset's mixed scenes. These structural irregularities highlight challenges for model training and evaluation and suggest future opportunities for applying specialised mitigation strategies, such as focal or class-balanced loss functions, synthetic data augmentation, or few-shot fine-tuning, to reduce bias towards dominant classes. Similarly, classes with a modest number of images but high object occurrence (Carrot, Cup) reflect densely annotated composite scenes, favouring models that can exploit contextual co-occurrence statistics. In contrast, mechanical and container objects (Mirror, Kettle, Cup, Bowl) exhibit splits closely aligned with the intended 75-80% guideline, providing robust validation sets despite their moderate frequency. For these classes, the primary challenge is not sample scarcity, but rather visual heterogeneity across lighting and pose—challenges that the cross-modal data capture strategy, particularly the depth and thermal modalities, is intended to mitigate. More broadly, the resulting class distribution reflects a long-tail structure, with notable variability in frequencies and a persistent imbalance between well-represented and rare categories. This composition encourages the development and evaluation of models that are robust to class imbalance and capable of generalising across a spectrum of representation levels. The MM5 split thus serves as a credible benchmark for multimodal fusion methods, supporting realistic performance assessment for rare-class generalisation, calibration, and uncertainty estimation, and inviting future research into imbalance-aware and few-shot learning paradigms.

4.5 Proposed Method

Our architecture leverages a dedicated stage-wise $\mathrm{RGB} + \mathrm{D} + \mathrm{I}$ fusion module, depicted in Figure 4.4, at each encoder scale to enrich RGB features. Crucially, the Intensity and Depth cues (along with surface Normals) provided to this module originate from our DIN modality. The DIN modality itself represents a data-level fusion, wherein

Table 4.2: Per-class training-evaluation split for the MM5 dataset's top level classes. Percentages are computed relative to the total number of images per class.

Class	Tot. Img.	Tot. Occ.	Train	Eval	Train %	Eval %
Apple	26	26	20	6	76.9	23.1
Apple Fake	26	26	19	7	73.1	26.9
Apple Green	47	53	36	11	76.6	23.4
Apple Green Bad	21	36	16	5	76.2	23.8
Apple Green Fake	20	20	15	5	75.0	25.0
Bowl	26	26	20	6	76.9	23.1
Carrot	24	51	18	6	75.0	25.0
Carrot Fake	18	27	14	4	77.8	22.2
Cup Cold	58	58	45	13	77.6	22.4
Cup Hot	20	20	16	4	80.0	20.0
Grapes Blue	17	17	14	3	82.4	17.6
Grapes Blue Bad	15	15	11	4	73.3	26.7
Grapes Blue Fake	16	16	13	3	81.2	18.8
Grapes Green	27	27	20	7	74.1	25.9
Grapes Green Bad	18	18	15	3	83.3	16.7
Grapes Green Fake	18	18	14	4	77.8	22.2
Kettle	9	9	7	2	77.8	22.2
Lemon	69	133	43	16	62.3	23.2
Lemon Bad	40	81	31	9	77.5	22.5
Lemon Fake	33	33	25	8	75.8	24.2
Lemon Half	24	44	18	6	75.0	25.0
Mandarin	47	133	34	13	72.3	27.7
Mandarin Bad	32	54	27	5	84.4	15.6
Mandarin Fake	21	21	17	4	81.0	19.0
Mandarin Half	17	20	14	3	82.4	17.6
Mandarin Peel	12	12	9	3	75.0	25.0
Mirror	29	29	20	9	69.0	31.0
Onion	21	30	16	5	76.2	23.8
Onion Red	21	33	16	5	76.2	23.8
Pear	24	30	18	6	75.0	25.0
Pear Bad	18	24	14	4	77.8	22.2

depth (D) and infrared intensity (I) images are geometrically pre-aligned with the RGB image. Surface normals (N), computed from the depth data, are included as a third channel, forming a compact D+I+N representation. By injecting features derived from this DIN modality into the RGB pathway at each stage, the network introduces complementary information related to surface structure (from D and N) and material properties (from I) early in the feature hierarchy, ensuring spatial co-registration with the RGB data. This strategy not only enhances the robustness of RGB features under challenging conditions, such as underexposure and overexposure, but also alleviates computational and representational demands on subsequent fusion stages involving additional modalities, including Thermal (T24) and UV. An overview of the overall architecture is depicted in Figure 4.2. The overall pipeline consists of: (i) a MiT-based encoder with stage-wise RGB+DIN enhancement (Section 4.5.1), (ii) a data-level fusion stage forming the DIN modality (Section 4.5.2), (iii) the SWIF module for gated cross-modal integration of Thermal and UV cues (Section 4.5.3), and (iv) a lightweight MLP decoder for multi-scale feature aggregation and prediction (Section 4.5.6). Thermal and UV modalities are integrated using learnable sigmoid gating mechanisms, inspired by established work in both sequence modelling and multimodal fusion [14, 15, 16, 17, 18, 19, 20]. At each encoder stage k, the thermal $(\mathrm{T24})$ and ultraviolet (UV) streams are first registered to the RGB pathway by the CM-FRM module [2]. Following feature extraction, patch embedding and transformer encoding, the aligned tensors $F^{(k)}_{\mathrm{RGB+DIN}}$, $F^{(k)}_{\mathrm{T24}}$, and $F^{(k)}_{\mathrm{UV}}$ are normalised and reshaped into spatial feature maps. Each auxiliary stream is processed by its own Sigmoid Gate module, following concepts introduced in the Gated Multimodal Unit (GMU) [17] and the Multimodal Transfer Module (MMTM) [19]. These modules consume the RGB+DIN base together with the aligned auxiliary feature map and output a pixel-wise confidence mask via a sigmoid activation, learning to weigh the auxiliary modality's contribution dynamically, conditioned on local content. Multiplying these masks with their respective auxiliary feature maps yields the gated contributions $G^{(k)}_{\mathrm{T24}}$ and $G^{(k)}_{\mathrm{UV}}$. The fused representation at each stage is thus given by:

$$F^{(k)}_{\mathrm{fused}} = F^{(k)}_{\mathrm{RGB+DIN}} + G^{(k)}_{\mathrm{T24}} + G^{(k)}_{\mathrm{UV}}. \tag{4.1}$$

This gating-based fusion scheme contrasts earlier approaches that use simple concatenation or static summation of features. Cheng et al. [18] pioneered a learnt gating approach for RGB-D semantic segmentation, in which a gate adaptively modulates the contribution of depth features. Similarly, Guo et al. [21] introduced DGFNet, applying dual gates to fuse spatially and semantically complementary information in land cover segmentation. Our approach generalises this principle, extending it to multiple auxiliary modalities with pixel-wise dynamic weighting. The theoretical foundations for gating originate from recurrent neural networks, notably the LSTM architecture by Hochreiter and Schmidhuber [14], which introduced sigmoid-activated gates to control the flow of information over time. This concept was extended to convolutional architectures by Dauphin et al. [15], who proposed Gated Linear Units (GLUs), leveraging sigmoid gates for selective information flow in deep CNNs. Channel-wise gating, as used in the Squeeze-and-Excitation (SE) module [16], brought adaptive recalibration of feature importance. In multimodal fusion, Arevalo et al. [17] introduced the Gated Multimodal Unit for soft weighting between modalities, and Joze et al. [19] proposed the Multimodal Transfer Module to facilitate learnable inter-modal transfer through gating. Most recently, Balit and Chadli [20] demonstrated the value of gated fusion in visible-thermal semantic segmentation. Our method integrates these advancements and, for the first time in a transformer-based segmentation architecture, applies learnable sigmoid gating at every fusion stage to achieve robust, context-adaptive integration of diverse modalities. By generating pixel-wise confidence masks conditioned on the primary RGB+DIN features, the network dynamically modulates the influence of auxiliary cues such as thermal and UV, activating them only when their information is contextually valuable. This fine-grained, stage-wise gating not only mitigates the propagation of modality-specific noise or sensor artefacts but also ensures that the fused representation remains optimally informative across varying scenes and conditions. As a result, the model demonstrates improved resilience to sensor failure and challenging imaging scenarios, outperforming traditional static or attention-based fusion strategies in both reliability and segmentation accuracy.

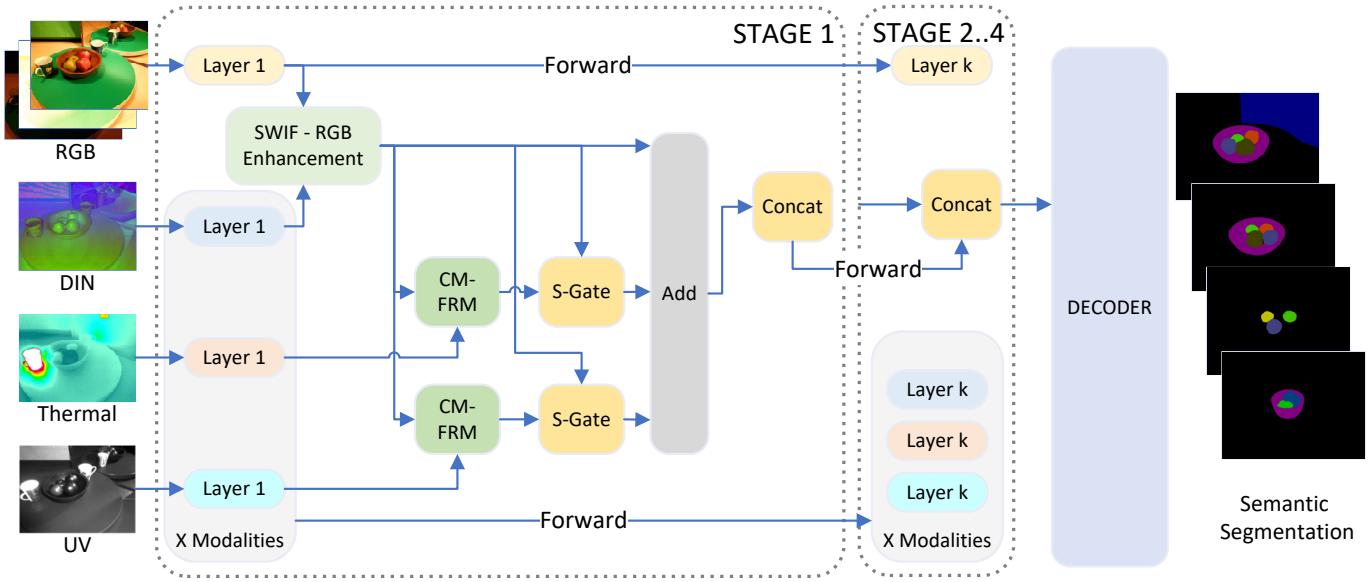


Figure 4.2: Encoder-decoder pipeline with stage-wise gated multimodal fusion. For each stage, modality features are linearly projected, concatenated across modalities, and passed through a lightweight gate generator to obtain per-pixel, per-modality gates, softly normalised before fusion. The fused representation at each stage is mapped to a common width, resized to quarter resolution, concatenated across stages, fused, and classified per pixel, followed by a final upsampling to full resolution. Stages 2-4 repeat the Stage 1 pipeline at progressively coarser scales, as indicated by the coloured "Layer 1" and "Layer k" blocks. See Equations (4.5)-(4.6) for the mathematical specification.

4.5.1 Encoder: Hierarchical MiT backbone with per-pixel gated multimodal fusion

We adopt a hierarchical Mix Transformer (MiT) backbone to encode each modality and produce four multi-scale feature maps that are fused per stage by learnt, per-pixel gates. Let the modality set be $\mathcal{M} = \{\mathrm{R}, \mathrm{D}, \mathrm{I}, \mathrm{T}, \mathrm{U}\}$ for RGB, depth, near-infrared intensity, thermal, and ultraviolet, respectively. For an input resolution (H, W), the MiT encoder yields stage features at $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$. We denote by $X_m^{(k)} \in \mathbb{R}^{B \times C_k \times H_k \times W_k}$ the stage-k feature

for modality $m \in \mathcal{M}$, where B is the batch size and C_k is the channel dimension at stage k, with the stage-k feature for modality $m \in \mathcal{M}$, with $(H_1, W_1) = (\frac{H}{4}, \frac{W}{4})$, $(H_2, W_2) = (\frac{H}{8}, \frac{W}{8})$, $(H_3, W_3) = (\frac{H}{16}, \frac{W}{16})$, and $(H_4, W_4) = (\frac{H}{32}, \frac{W}{32})$. Figure 4.2 illustrates the complete architecture, showing how the tensors $X_m^{(k)}$ are projected, gated, fused into $F_{\text{fused}}^{(k)}$, and routed to the decoder.

Modality-wise normalisation Before encoding, each modality is standardised per channel. Two variants are considered: (i) default normalisation using ImageNet statistics, with mean [0.485,0.456,0.406] and standard deviation [0.229,0.224,0.225] in the [0,1] range applied to all modalities, and (ii) specific normalisation, where dataset-wide statistics are calculated separately for each channel of each modality. The latter ensures that channels with very different native value ranges, for example thermal versus ultraviolet, are brought to a comparable numerical scale. This stabilises optimisation and prevents the gating modules from responding merely to amplitude differences across modalities rather than to informative content. Formally, we pre-compute a mean and a standard deviation for every channel c of every modality m by aggregating all pixels from all images of that modality in the dataset:

$$\mu_{m,c} \;=\; \frac{1}{N_{m,c}} \sum_{p=1}^{N_{m,c}} \frac{x_p^{(m,c)}}{255}, \qquad \sigma_{m,c} \;=\; \sqrt{\frac{1}{N_{m,c}} \sum_{p=1}^{N_{m,c}} \left(\frac{x_p^{(m,c)}}{255}\right)^2 - \mu_{m,c}^2}, \tag{4.2}$$

where $N_{m,c}$ is the total number of pixels across all images for modality m and channel c. At training and inference time, each incoming image channel, written as $I^{(m,c)} \in [0, 1]$ after division by 255, is standardised by subtracting the corresponding dataset mean and dividing by the dataset standard deviation,

$$\bar{I}^{(m,c)} \;=\; \frac{I^{(m,c)} - \mu_{m,c}}{\sigma_{m,c} + \varepsilon}, \tag{4.3}$$

with a small constant ε added to avoid division by zero. In practice, the dataset statistics used in Equation (4.2) are computed offline by a script that scans each modality folder, accumulates sums and sums of squares per channel, and outputs the resulting $\mu_{m,c}$ and $\sigma_{m,c}$ in the $[0, 1]$ domain for direct use in the data loader. The empirical impact of specific normalisation on optimisation stability and validation accuracy is analysed in Figure 4.8.
To evaluate the impact of normalisation schemes, we employ three deterministic seed-augmentation configurations for controlled experiments. Each configuration is used for both normalisation schemes to enable direct comparisons. The configurations are: Set 1 with scale 1.00, angle 0°; Set 2 with scale 0.95, angle +5°; and Set 3 with scale 1.05, angle $-5°$. All three runs are trained for the same number of epochs with identical hyperparameters, differing only in the random seed and augmentation parameters specified above. This setup enables us to compute the mean performance with 95% confidence intervals and conduct matched statistical analysis between normalisation schemes. The empirical results of this comparison are presented in Section 4.5.7.

Stage-wise projections For each stage k, the output feature map from each modality is aligned to a common channel dimension C_k using a pointwise linear projection, implemented as a 1×1 convolution. This operation transforms the channel dimension while preserving the spatial dimensions (H_k, W_k) unchanged:

$$\hat{X}_m^{(k)} \;=\; \text{Linear}_{k,m}\Big(C_{m,\text{in}}^{(k)} \to C_k\Big)\Big(X_m^{(k)}\Big), \quad m \in \mathcal{M}. \tag{4.4}$$

where $C_{m,\text{in}}^{(k)}$ denotes the input channel dimension for modality m at stage k (which varies across modalities), and C_k is the target common channel dimension for fusion at stage k as specified below.

Per-pixel sigmoid gating and fusion At each stage k, auxiliary modalities are integrated through independent sigmoid gating. After the SWIF module produces $F_{\text{RGB+DIN}}^{(k)}$ (the RGB features enhanced with DIN), each auxiliary modality undergoes individual processing. For thermal (T) and ultraviolet (UV) streams:

1. The CM-FRM module aligns each auxiliary modality to $F_{\text{RGB+DIN}}^{(k)}$
2. Each aligned auxiliary modality passes through its own Sigmoid Gate module

The Sigmoid Gate module for modality m takes as input the concatenation of the base features and the aligned auxiliary features:

$$G_m^{(k)} = \sigma\left(\phi_m^{(k)}\left(\text{Concat}\left(F_{\text{RGB+DIN}}^{(k)}, F_m^{(k)}\right)\right)\right) \odot F_m^{(k)} \tag{4.5}$$

where $\phi_m^{(k)}$ is a per-pixel MLP that produces a single-channel spatial attention map, and σ is the sigmoid activation. The gated contributions from T and UV are then added to the base features:

$$F_{\text{fused}}^{(k)} = F_{\text{RGB+DIN}}^{(k)} + G_{\text{T}}^{(k)} + G_{\text{UV}}^{(k)} \tag{4.6}$$

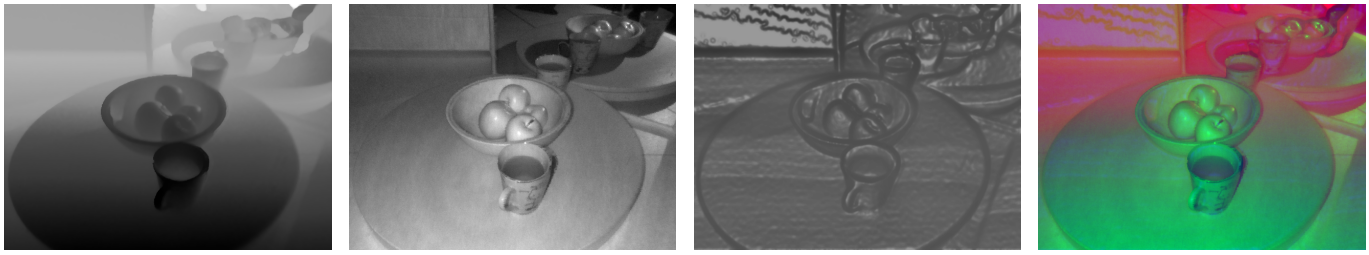
This per-pixel gating mechanism allows the network to adaptively weight each auxiliary modality's contribution based on local content, suppressing unreliable or uninformative signals where needed. Stages 2 through 4 repeat this pipeline with stage-specific parameters. The set of fused features $\{F_{\text{fused}}^{(k)}\}_{k=1}^{4}$ from all four stages forms the multi-scale output of the encoder and is forwarded to the decoder head described in Section 4.5.6.

Channel widths Unless otherwise noted, we adopt canonical MiT settings for the common fusion dimensions: for MiT-B0, $(C_1, C_2, C_3, C_4) = (32, 64, 160, 256)$; for MiT-B2, $(C_1, C_2, C_3, C_4) = (64, 128, 320, 512)$. These values determine the channel dimensions after projection: each $\hat{X}_m^{(k)} \in \mathbb{R}^{B\times C_k\times H_k\times W_k}$ after the linear projection in Equation 4.4, and consequently the fused outputs $F_{\text{fused}}^{(k)} \in \mathbb{R}^{B\times C_k\times H_k\times W_k}$ consumed by the decoder.

Mapping into the decoder Each fused stage output $F_{\text{fused}}^{(k)}$ has dimension C_k and spatial resolution (H_k, W_k) where $(H_1, W_1) = (\frac{H}{4}, \frac{W}{4})$, $(H_2, W_2) = (\frac{H}{8}, \frac{W}{8})$, $(H_3, W_3) = (\frac{H}{16}, \frac{W}{16})$, and $(H_4, W_4) = (\frac{H}{32}, \frac{W}{32})$. Each is mapped by a 1×1 layer to a common decoder width C and resized to $\frac{H}{4} \times \frac{W}{4}$. The four resized tensors are concatenated along channels to $\mathbb{R}^{B\times 4C\times \frac{H}{4}\times\frac{W}{4}}$, fused by a second 1×1 layer back to $\mathbb{R}^{B\times C\times \frac{H}{4}\times\frac{W}{4}}$, and classified per pixel to $\mathbb{R}^{B\times N_{\text{cls}}\times \frac{H}{4}\times\frac{W}{4}}$. A final bilinear upsampling by $\times4$ produces $\hat{Y} \in \mathbb{R}^{B\times N_{\text{cls}}\times H\times W}$.

4.5.2 Data-level Fusion

Depth-Intensity-Normal (DIN) Composite. We first fuse depth and infrared intensity at the data level by augmenting the depth map with its associated surface normals as shown in Figure 4.3. This "DIN" composite, constructed from the aligned depth, intensity, and normal channels, is injected into the encoder at each stage via a residual enhancement block. The normals are derived from depth gradients and smoothed to ensure spatial coherence; the full preprocessing pipeline is detailed in 4.B.1.



(a) Depth image (b) Infrared intensity (c) Surface normals (d) Merged DIN

Figure 4.3: Data-level Depth-Intensity-Normals fusion: (a) depth image, (b) infrared intensity image, (c) surface normals computed from depth, (d) merged depth-intensity (DIN) representation.

4.5.3 Stage-Wise Intensity Fusion(SWIF)

In multimodal computer vision, effectively leveraging complementary cues from both colour and geometric information is central to robust scene understanding, especially under challenging illumination conditions. While RGB features provide rich semantic and appearance detail, they are inherently sensitive to lighting variation, shadows, and saturation effects. In contrast, intensity and depth-derived features offer illumination-invariant cues about object

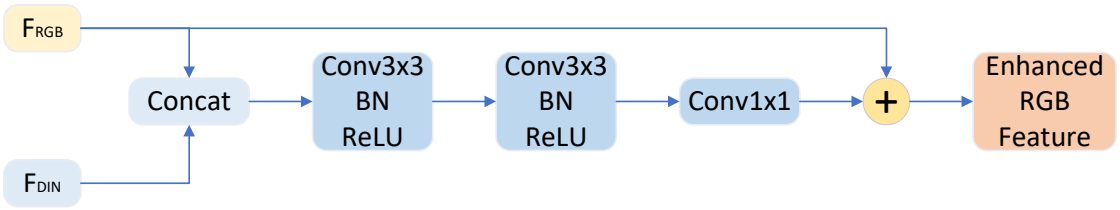


Figure 4.4: SWIF (Stage-Wise Intensity Fusion) Module: The RGB and intensity feature maps are concatenated channel-wise and passed through a three-layer convolutional fusion network, producing an enhancement which is summed with the original RGB features via a residual connection to yield the stage-wise enhanced RGB representation.

boundaries and surface structure but may lack fine-grained textural information. Simply fusing these modalities at a late network stage often limits the model's ability to capture subtle, local correlations between colour, shape, and luminance. To address this, we introduce a stage-wise RGB-DIN enhancement module, SWIF, that injects complementary intensity and depth information directly into the RGB feature stream at every encoder depth. This early, spatially co-registered fusion enables the network to learn meaningful interactions between modalities throughout the feature hierarchy, resulting in enhanced feature representations that are both structurally aware and resilient to lighting artefacts. By integrating these cues before high-level fusion or decoding, the network can emphasise illumination-invariant edges and object boundaries within the learnt colour features, ultimately improving segmentation accuracy and convergence. The processing steps executed by this module are outlined below, and an overview is presented in Figure 4.4. At stage k, we denote the RGB and DIN feature tensors by

$$F_{\text{RGB}}^{(k)} \in \mathbb{R}^{B\times C\times H\times W}, \qquad F_{\text{DIN}}^{(k)} \in \mathbb{R}^{B\times C\times H\times W}, \tag{4.7}$$

where B is batch size, C the channel count, and H, W the spatial dimensions. Concatenating the two maps along the channel axis gives

$$X^{(k)} = [F_{\text{RGB}}^{(k)}\,;\, F_{\text{DIN}}^{(k)}] \in \mathbb{R}^{B\times 2C\times H\times W}, \tag{4.8}$$

which is processed by a compact fully convolutional sub-network,

$$\Phi^{(k)} : \mathbb{R}^{B\times 2C\times H\times W} \longrightarrow \mathbb{R}^{B\times C\times H\times W}. \tag{4.9}$$

The sub-network comprises a 3×3 convolution that reduces channels from $2C$ to C_{hid}, followed by batch normalisation and ReLU; a second 3×3 convolution that preserves C_{hid} channels, again followed by batch normalisation and ReLU; and a final 1×1 convolution that restores the width to C. The hidden width is

$$C_{\text{hid}} = \max(16, \lfloor C\cdot r \rfloor), \qquad r = 0.5 \text{ by default.} \tag{4.10}$$

The enhancement generated at this depth is

$$E^{(k)} = \Phi^{(k)}(X^{(k)}), \tag{4.11}$$

and the refined RGB+DIN base for subsequent fusion is obtained through a residual addition,

$$F_{\text{RGB+DIN}}^{(k)} = F_{\text{RGB}}^{(k)} + E^{(k)}. \tag{4.12}$$

All convolution and normalisation layers are initialised with Kaiming [22] style schemes: convolutional kernels are sampled from $\mathcal{N}(0, \sqrt{2/\text{fan\_out}})$, while the scale parameters of the batch normalisation layers are set to one with zero offset, ensuring stable end to end optimisation from the outset.

4.5.4 Auxiliary Modality Alignment

At each encoder stage, the thermal and ultraviolet feature maps are spatially registered to the main pathway using the Cross-Modal Feature Rectification Module (CM-FRM) [2]. The CM-FRM, initially proposed in the CMX architecture, rectifies each modality's features by adaptively combining spatial and channel-wise information from both

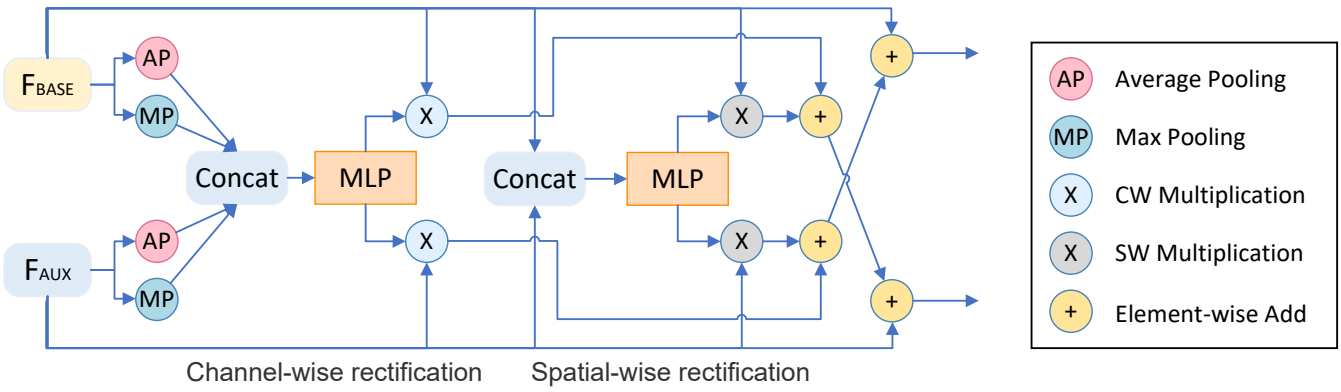


Figure 4.5: The CM-FRM (Cross-Modal Feature Rectification Module [2]) architecture, detailing its channel-wise and spatial-wise feature refinement pathways for multimodal inputs.

the reference (here, RGB+DIN) and the auxiliary stream. In practice, CM-FRM leverages convolutional attention and affine transformation to align features robustly, mitigating parallax and resolution differences between modalities [2]. Unlike classical parametric transformations, such as thin-plate spline (TPS) [23, 24] registration, which learns a global warping function based on control points, the CM-FRM enables local, feature-level adaptation within the deep network and is less susceptible to overfitting or producing artefacts when modality-specific distortions are present. An overview of the module is shown in Figure 4.5.
While we evaluated TPS-based spatial transformers for feature alignment, we observed that the CM-FRM provided superior performance in terms of both stability and segmentation accuracy on the MM5 dataset. The TPS approach, although powerful for modelling smooth, global deformations, was less effective for local, fine-grained rectification.

4.5.5 Auxiliary Modality Fusion

The effective integration of auxiliary modalities with a primary data stream, such as RGB imagery, can significantly enhance perception tasks by providing complementary information that is not always accessible in the visible spectrum. For adaptive fusion, where the contributions of auxiliary sources are dynamically modulated based on the input context, gated mechanisms are particularly effective, as they enable fine-grained control over feature propagation. In our framework, sigmoid gating leverages the output range of the sigmoid function, $[0, 1]$, to generate learnable, pixel-wise weights, also known as 'confidence masks'. These masks enable the network to selectively regulate the influence of each auxiliary modality, thereby highlighting salient features and suppressing irrelevant or noisy signals. This approach is crucial for robust multimodal fusion, especially when auxiliary sensors may experience modality-specific artefacts or unreliable readings. Our auxiliary modality fusion module performs stage-wise integration of thermal (T) and ultraviolet (UV) features into the main representation using independent, context-aware sigmoid gating. At each encoder stage k, the T and UV streams are first passed through patch embedding and transformer encoding blocks in parallel to the main RGB(+DIN) stream. The CM-FRM module aligns the feature representation of each auxiliary modality to the base feature representation. After transformer encoding, the outputs $F^{(k)}_{\text{RGB+DIN}}$, $F^{(k)}_{\text{T}}$, and $F^{(k)}_{\text{UV}}$ are passed through layer normalisation, as per standard transformer architecture, and reshaped into spatial feature maps. The thermal and UV features are then modulated by their respective gates, which compute a spatially varying contribution based on the concatenated local context of the base and auxiliary features. The gated auxiliary contributions are added to the base feature map to produce the fused representation at each stage. For each auxiliary stream, a dedicated Sigmoid Gate module receives the concatenation of the base and aligned auxiliary feature maps. The gate generation pathway consists of a lightweight MLP followed by a sigmoid activation, producing a single-channel, spatially varying mask. Simultaneously, the auxiliary feature map is transformed (via a 1×1 convolution or identity mapping) and multiplied element-wise by the generated gate, producing the gated auxiliary contribution $G^{(k)}_{\text{aux}}$. This process is performed independently for both thermal and UV branches, resulting in $G^{(k)}_{\text{T}}$ and $G^{(k)}_{\text{UV}}$. The

fused representation at each stage is then computed by summing the base map and all available gated contributions:

$$F_{\text{fused}}^{(k)} = F_{\text{RGB+DIN}}^{(k)} + \sum_{i \in \{\text{T, UV}\}} G_i^{(k)}. \tag{4.13}$$

This flexible, adaptive gating strategy enables the network to selectively incorporate auxiliary information where and when it is most beneficial, and to disregard unhelpful or corrupted signals, although the gates learn their behaviour in training and are fixed at inference. The overall structure and processing flow of the sigmoid gate module are illustrated in Figure 4.6, which provides an overview of the key operations involved in context-aware gating and fusion.

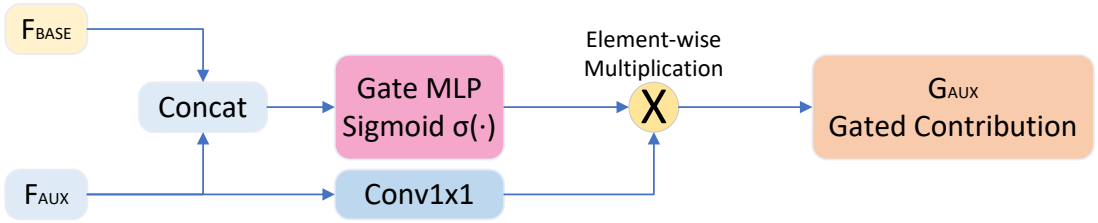


Figure 4.6: Sigmoid Gate Module: The module takes a base feature map (F_{base}) and an auxiliary modality feature map (F_{aux}). For gate generation, F_{base} and F_{aux} are concatenated and processed by a Gate MLP followed by a Sigmoid function to produce a 1-channel gate. Separately, F_{aux} is processed by a Transform Conv (a 1×1 convolution or an identity operation) to produce the transformed auxiliary feature. This transformed feature is then element-wise multiplied by the gate to yield the gated contribution G_{aux}.

4.5.6 Decoder Architecture

The set of fused feature maps is forwarded to a lightweight MLP decoder head that aggregates multi-scale cues into dense predictions. Following SegFormer's design, as adopted by CMX, the decoder uses only pointwise linear projections and bilinear interpolation, which keeps memory and latency low while effectively mixing information across scales [25, 2]. We retain this canonical decoder unchanged to keep it lightweight and to isolate the contribution of our gated encoder from decoder modifications.

Let the encoder yield $\{F_{\text{fused}}^{(k)}\}_{k=1}^{4}$ with $F_{\text{fused}}^{(k)} \in \mathbb{R}^{B \times C_k \times H_k \times W_k}$ at resolutions $(H_1, W_1) = (\frac{H}{4}, \frac{W}{4})$, $(H_2, W_2) = (\frac{H}{8}, \frac{W}{8})$, $(H_3, W_3) = (\frac{H}{16}, \frac{W}{16})$, $(H_4, W_4) = (\frac{H}{32}, \frac{W}{32})$. Each map is projected to a common width C by a pointwise linear layer, implemented as a 1×1 convolution, then resized to quarter resolution:

$$\hat{F}^{(k)} = \text{Linear}(C_k \to C)\left(F_{\text{fused}}^{(k)}\right), \qquad \tilde{F}^{(k)} = \text{Upsample}\left(\tfrac{H}{4}, \tfrac{W}{4}\right)\left(\hat{F}^{(k)}\right). \tag{4.14}$$

The resized maps are concatenated along channels and fused by a second linear mapping:

$$F_{\text{cat}} = \text{Concat}\left(\tilde{F}^{(1)}, \tilde{F}^{(2)}, \tilde{F}^{(3)}, \tilde{F}^{(4)}\right) \in \mathbb{R}^{B \times 4C \times \frac{H}{4} \times \frac{W}{4}}, \qquad F_{\text{fuse}} = \text{Linear}(4C \to C)(F_{\text{cat}}). \tag{4.15}$$

A per-pixel classifier produces quarter-resolution logits:

$$M = \text{Linear}(C \to N_{\text{cls}})(F_{\text{fuse}}) \in \mathbb{R}^{B \times N_{\text{cls}} \times \frac{H}{4} \times \frac{W}{4}}. \tag{4.16}$$

Finally, the logits are upsampled by a factor of four to the input resolution:

$$\hat{Y} = \text{Upsample}(\times 4)(M) \in \mathbb{R}^{B \times N_{\text{cls}} \times H \times W}. \tag{4.17}$$

This decoder cleanly separates fusion in the encoder from prediction in the head, leveraging the complementary strengths of low-level, high-resolution features together with semantically rich, low-resolution features. Unlike architectures for RGB-D-T salient object detection that may forward only a subset of encoder stages, our segmentation

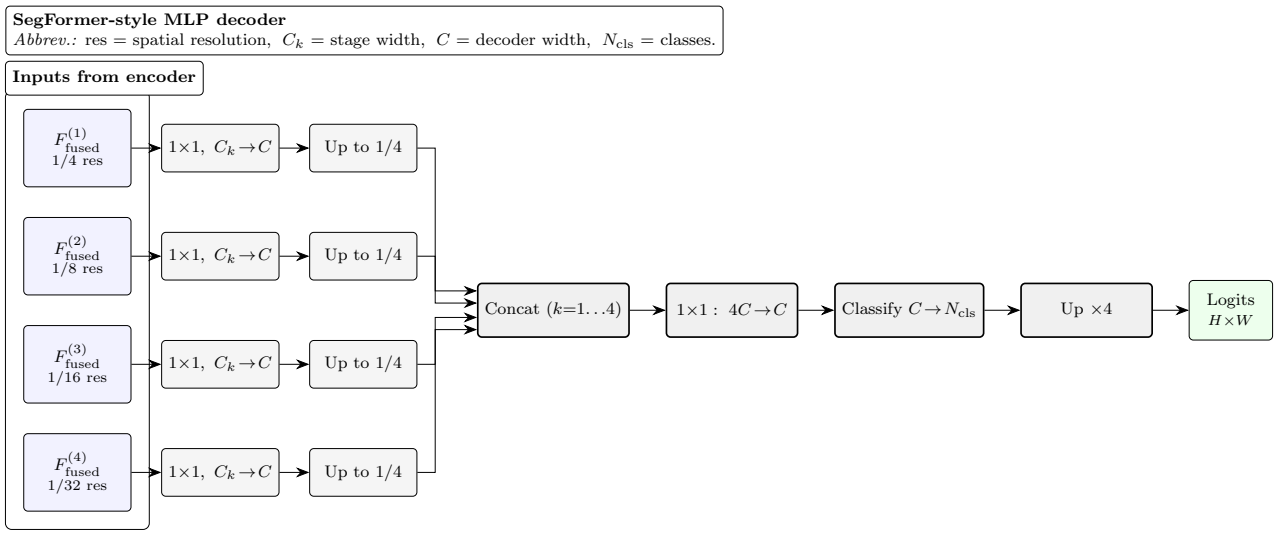


Figure 4.7: MLP decoder schematic. Each fused encoder feature $F^{(k)}_{\text{fused}}$ is linearly projected to a common width C, resized to quarter resolution, concatenated across scales, fused by a pointwise linear layer and classified per pixel, then upsampled $\times 4$ to full resolution. The design contains no attention blocks or deconvolutions, following the SegFormer and CMX practice.

decoder consumes all four fused scales to preserve both fine detail and global semantics. Figure 4.7 visualises this pipeline, showing the per-stage projection, resizing, concatenation, fusion, and final upsampling operations.

4.5.7 Training Procedure

This section outlines the empirical methodology employed for model training and evaluation, detailing the computational environment, parameter initialisation protocols, optimisation strategies, data handling techniques, and the rationale behind backbone-specific training schedules. These procedures are designed to ensure reproducibility and are grounded in established deep learning practices.

Experimental Setup, Parameter Initialisation, and Optimisation

All experiments were conducted using the PyTorch deep learning framework. Computations were performed on a workstation equipped with an Intel Core i7-13700F CPU and a single NVIDIA RTX 3090 GPU. This hardware configuration influenced choices such as mini-batch sizing, particularly for models with substantial parameter counts (e.g., MiT-B2), thereby underscoring the importance of computationally efficient training strategies. Parameter initialisation was consistent across all model components. Weights for both Mix Transformer (MiT) backbones, specifically MiT-B0 and MiT-B2, based on the SegFormer architecture [25], along with all newly introduced fusion and gating blocks, were initialised from scratch. The MiT-B1 variant was not included in our evaluation, as its architecture and computational profile are strictly intermediate between B0 and B2, differing only in embedding size and not in qualitative design. Prior work [25] has demonstrated that MiT-B1's empirical performance predictably interpolates between that of B0 and B2, without offering unique insights. This approach, eschewing pre-trained weights, renders the training process more sensitive to the characteristics of the training dataset and the duration of training, potentially accentuating phenomena such as epoch-wise double descent, especially for larger models. A Kaiming normal distribution was employed to initialise all network weights, a standard practice for architectures employing ReLU-like activation functions, to mitigate issues of vanishing or exploding gradients. All biases throughout the network were uniformly initialised to zero. The AdamW optimiser [26] was selected for its efficacy in training deep neural networks, particularly transformer-based architectures, due to its improved handling of weight decay by decoupling it from the adaptive learning rate mechanism. The optimiser was configured with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. An initial learning rate of 1×10^{-3} was set, complemented by a weight decay coefficient of 10^{-2} to provide regularisation. The learning rate schedule incorporated a linear warm-up phase spanning the first ten epochs, which aids in stabilising training during the initial stages, especially when using relatively large learning

rates. Following the warm-up, the learning rate was subjected to a polynomial decay with a power of 0.9, a typical schedule for gradually annealing the learning rate towards the end of training. A mini-batch size of 8 was consistently used for all experiments. While potentially constrained by GPU memory capacity for larger models, this batch size influences the stochasticity of the gradient estimates and overall training dynamics. During training, each iteration involved sampling an image from the MM5 dataset, along with its associated auxiliary inputs: depth-intensity-normal (DIN) fusion and DTMRE-encoded thermal T24 channels [1]. To standardise the input data, each channel was independently normalised to have zero mean and unit variance, using precomputed statistics from the MM5 dataset. Data augmentation techniques were applied to enhance model generalisation, including random horizontal flipping and multi-scale resizing. Optimisation was driven by a composite CEDice loss function, wherein the cross-entropy (CE) and Dice loss components were accorded equal weighting (i.e., $0.5 \times \text{CE} + 0.5 \times \text{Dice}$). This composite loss structure effectively balances pixel-wise classification accuracy, derived from the CE term, with considerations of volumetric overlap from the Dice term, which is particularly advantageous for semantic segmentation tasks, especially in the presence of class imbalance. Manual class weights were applied to address class imbalance and focus the model's learning capacity on foreground classes of interest, reducing the influence of the typically prevalent background class to 0.1.

Capacity-Data Trade-offs and Regularisation Strategies

The training dynamics of the MiT-B0 and MiT-B2 backbones reveal distinct interactions between model capacity, training duration, and regularisation requirements. The model configured with the MiT-B2 backbone, comprising a total of approximately 140 million parameters, achieved a validation mean Intersection over Union (mIoU) of 86.1% after 250 training epochs. Extending training to 500 epochs resulted in a marginal improvement to 86.5%, indicating diminishing returns for prolonged training of high-capacity models. This plateau suggests that, beyond a certain point, additional epochs may not substantially benefit such models, and emphasis should instead be placed on effective regularisation techniques. These may include data augmentation, dropout, and weight decay, which have been shown to mitigate overfitting in large neural networks [27, 28]. In contrast, the MiT-B0 backbone, with approximately 24 million parameters, demonstrated significant improvements with extended training. Training for 500 epochs increased the validation mIoU from 86.2% at 250 epochs to 88.3%, outperforming the MiT-B2 model trained for the same duration by 1.8 percentage points. This suggests that smaller models benefit from longer training schedules, allowing them to better explore the loss landscape and achieve improved generalisation [27, 28].

Training dynamics under different normalisation schemes

We compare the default ImageNet normalisation with the dataset-specific modality-wise normalisation defined in Equations (4.2) and (4.3) under identical settings: the same backbone, optimiser, data splits, and static augmentations. To assess reproducibility, we use three deterministic configurations, referred to as Set 1, Set 2, and Set 3, each with a fixed seed and fixed augmentation parameters. Set 1 has static mirroring, scale factor 1.00, and rotation 0°; Set 2 has static mirroring, scale factor 0.95, and rotation +5°; Set 3 has static mirroring, scale factor 1.05, and rotation $-5°$. Each set is trained once with default normalisation and once with dataset-specific normalisation.

Averaged across the three sets, the dataset-specific scheme attains a mean final validation mIoU of 0.8005 with 95% confidence interval [0.7456, 0.8555], compared with 0.7816 with 95% confidence interval [0.7242, 0.8389] for the default scheme. The mean paired improvement is 0.0190 (95% CI: [-0.0125, 0.0505]), representing a 2.4% relative improvement. In terms of convergence speed, the dataset-specific scheme demonstrates substantial acceleration: median epochs to reach 0.75, 0.78, and 0.81 mIoU are 14, 17, and 30, versus 17, 24, and 47 for the default scheme—reductions of 18%, 29%, and 36% respectively.

Figure 4.8 shows the mean training loss and the mean validation mIoU across the three sets with one-standard-deviation bands. The dataset-specific normalisation yields consistently faster loss reduction and more stable convergence to higher validation accuracy. This makes it a sound training choice that improves optimisation behaviour and validation accuracy at no inference cost.

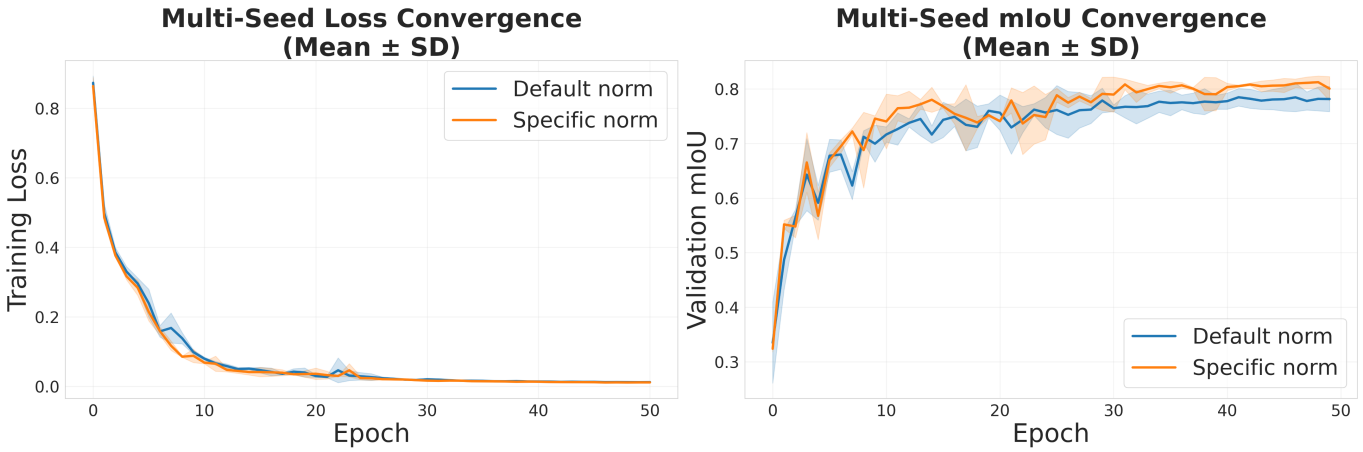


Figure 4.8: Effect of normalisation on training dynamics under identical settings, with one-standard-deviation bands across three deterministic configurations. Left: mean training loss. Right: mean validation mIoU. The dataset-specific scheme shows faster convergence and improved validation accuracy.

4.6 Results and Discussion

The semantic segmentation performance was evaluated across various input modality combinations and under different RGB lighting conditions: ideal ('RGB3'), underexposed ('RGB1'), and overexposed ('RGB5'). The core modalities include the data-level fused depth, intensity, and normals ('DIN'), a preprocessed thermal stream ('T24') designed to preserve minute temperature differences, and an ultraviolet stream ('U8'). Classes denoted as 'bad' refer to partially rotten fruit exhibiting distinct thermal signatures, while 'fake' classes are plastic replicas. We analyse the mean Intersection over Union (mIoU) and pixel accuracy. The detailed IoU scores for each class are presented in Table 4.6 using a MiT-B0 and Table 4.7 using a MiT-B2, while an overview of the overall results is presented in Table 4.3.

4.6.1 Evaluation Metrics and Comparative Analysis

To ensure a thorough and objective comparison of model performance across all categories, we report a suite of widely adopted evaluation metrics.

Mean Intersection over Union (*Mean IoU*): This metric is obtained by calculating the Intersection over Union (IoU) for each class, defined as the ratio of the overlap between predicted and ground-truth regions to their union, and then averaging these values across all classes. Mean IoU offers a class-balanced measure of overall segmentation accuracy.

Frequency Weighted Intersection over Union (*Freq IoU*): Here, the IoU for each class is weighted according to its frequency in the dataset, thereby aligning the metric with the dataset's inherent class distribution. This approach places greater emphasis on the performance of prevalent classes.

Mean Pixel Accuracy (*Mean Pixel Acc*): Mean Pixel Accuracy is computed as the average of per-class pixel accuracies, where pixel accuracy reflects the fraction of correctly classified pixels for a given class. This metric is sensitive to performance across both frequent and rare classes.

Pixel Accuracy (*Pixel Acc*): This measures the proportion of all pixels in the dataset that are classified correctly, irrespective of their class labels, providing a straightforward indicator of global segmentation performance.

Mean Rank: To facilitate equitable comparison between models, we also report the *Mean Rank* for each method [29]. Within each class, models are ranked according to their performance, with rank 1 assigned to the best performing method, and ties receiving an average rank. The mean rank of each model is then calculated as the average of its class-wise ranks, offering an interpretable, class-balanced summary of comparative performance across the full class set.

4.6.2 Overall Performance

The primary metric for inter-class comparison is the mean Intersection over Union (mIoU). The highest overall performance, with a mean IoU of 88.3%, was achieved with the four-stream combination of ideal RGB, DIN,

Table 4.3: Detailed IoU and network statistics for various modality combinations and lighting conditions, for both MiT-B0 (top) and MiT-B2 (bottom) backbones. DIN: Depth-Intensity-Normals fused; T24: processed thermal; U8: ultraviolet; RGB1: under-exposed RGB; RGB3: ideal RGB; RGB5: over-exposed RGB. "Bad" classes are partially rotten; "Fake" classes are replicas.

Class	2 RGB1-U8	2 RGB1-T24	2 RGB3-DIN	3 RGB1-DIN-U8	3 RGB3-DIN-U8	3 RGB5-DIN-U8	3 RGB1-DIN-T24	3 RGB3-DIN-T24	3 RGB5-DIN-T24	4 RGB1-DIN-T24-U8	4 RGB3-DIN-T24-U8	4 RGB5-DIN-T24-U8
MiT-B0 (500 epochs)												
Mean IoU	58.3	60.1	80.7	72.5	81.9	73.0	85.6	86.6	84.7	84.9	**88.3**	84.2
Freq IoU	98.6	98.6	99.3	99.1	99.4	99.1	99.4	99.5	99.5	99.4	**99.6**	99.4
Mean Pixel Acc	71.1	72.4	88.7	82.5	88.9	82.5	92.3	92.3	91.3	92.4	**93.9**	91.1
Pixel Acc	99.1	99.1	99.6	99.5	99.7	99.5	99.7	**99.8**	99.7	99.7	**99.8**	99.7
Mean Rank	11.1	10.4	6.7	9.0	5.3	8.4	5.4	3.7	5.4	5.6	**2.2**	4.8
FPS	104	104	104	74	74	74	74	74	74	55	55	55
Parameters	11M	11M	11M	18M	18M	18M	18M	18M	18M	24M	24M	24M
GFLOPs	10.8	10.8	10.8	14.5	14.5	14.5	14.5	14.5	14.5	17.3	17.3	17.3
MiT-B2 (250 epochs)												
Mean IoU	45.8	58.2	75.3	73.0	78.3	75.7	82.4	78.9	81.6	83.8	**86.1**	83.6
Freq IoU	93.8	98.5	99.0	99.1	99.3	99.2	99.1	99.1	99.3	99.3	**99.5**	99.3
Mean Pixel Acc	63.9	69.7	85.5	83.7	86.9	85.7	91.8	87.4	90.3	92.4	**93.7**	91.9
Pixel Acc	95.8	99.1	99.4	99.5	99.6	99.5	99.5	99.5	99.6	99.6	**99.7**	99.6
Mean Rank	10.7	10.5	6.4	9.0	5.8	7.8	5.7	3.3	6.3	4.8	**2.9**	4.9
FPS	39	39	39	29	29	29	29	29	29	25	25	25
Parameters	67M	67M	67M	106M	106M	106M	106M	106M	106M	140M	140M	140M
GFLOPs	60.9	60.9	60.9	84.8	84.8	84.8	84.8	84.8	84.8	105.0	105.0	105.0

thermal, and ultraviolet ('RGB3-DIN-T24-U8'). Even under suboptimal conditions, the network remains robust: with underexposed RGB ('RGB1-DIN-T24-U8'), a mean IoU of 84.9% is achieved, and with overexposed RGB ('RGB5-DIN-T24-U8'), the performance is still 84.2%. These results demonstrate the substantial benefit of fusing diverse sensor streams for reliable segmentation under varying lighting. Further, the frequency-weighted IoU and pixel accuracy for the best four-stream configuration reach 99.6% and achieve the best mean rank of 2.2 across all evaluated combinations, highlighting both accuracy and consistency. Performance gains from multimodal fusion are not limited to ideal lighting: for example, adding DIN to underexposed RGB boosts mean IoU from 60.1% (RGB1-T24) to 85.6% (RGB1-DIN-T24), confirming the critical contribution of geometrically aligned depth and intensity features. The choice of backbone architecture also shapes the trade-off between accuracy and efficiency. The MiT-B0 model delivers a favourable balance between segmentation accuracy and computational efficiency, outperforming the heavier MiT-B2 backbone in both speed and mean IoU for the four-stream setup. The MiT-B0 backbone in the largest fusion setting uses 24 million parameters and 17.3 GFLOPs, whereas the MiT-B2 backbone requires 140 million parameters and 105 GFLOPs, with only a marginal change in mean IoU. With the available dataset size, the larger and more computationally demanding MiT-B2 backbone did not yield accuracy gains over MiT-B0. Thus, MiT-B0 provides the most practical solution, combining high segmentation accuracy with low resource requirements and real-time performance. A sample of predictions and associated input images is shown in Figure 4.9.

4.6.3 Impact of Lighting Conditions

The quality of the RGB input had a significant influence on overall performance, although the multimodal setup provided considerable resilience.

- **Ideal Lighting ('RGB3'):** Configurations with 'RGB3' consistently produced the best results within their respective modality groups. For instance, 'RGB3-DIN-T24' (mean IoU 86.6%) outperformed 'RGB1-DIN-T24' (mean IoU 85.6%) and 'RGB5-DIN-T24' (mean IoU 84.7%). The combination 'RGB3-DIN-T24-U8' yielded the top mean IoU of 88.3%.

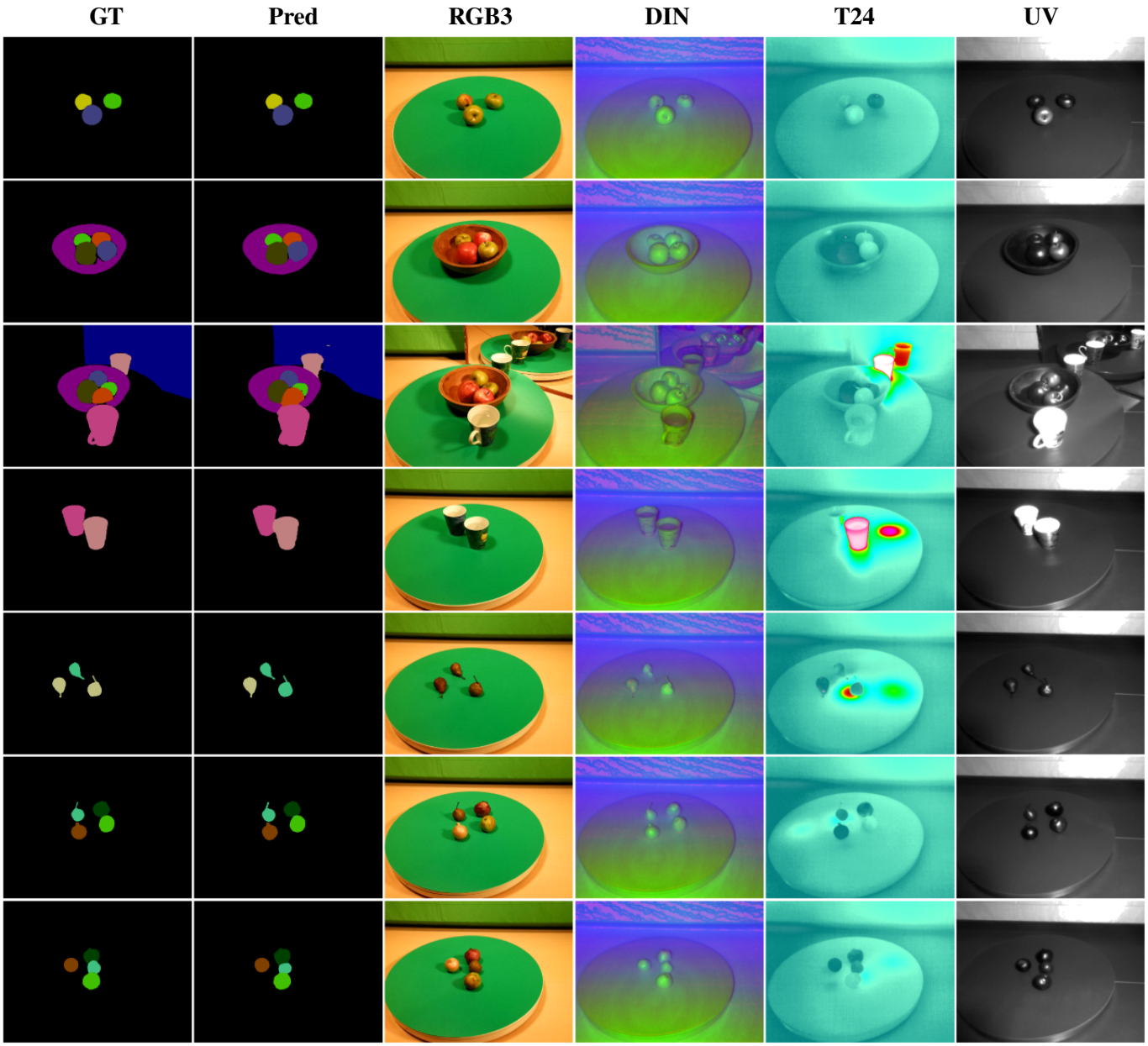


Figure 4.9: Example multimodal segmentation results for seven selected frames: from left to right, ground truth mask, predicted mask, RGB, DIN (depth-intensity-normals), thermal (T24), and ultraviolet (UV). Each row corresponds to a different frame (top to bottom: 240, 250, 256, 263, 271, 289, 294).

- **Underexposed RGB ('RGB1'):** The system demonstrated substantial robustness to underexposure, with the 'RGB1-DIN-T24-U8' configuration achieving a mean IoU of 84.9%. Although this falls short of the ideal-light counterpart (88.3%), it represents a marked improvement over single- or dual-modality variants under low-light conditions (e.g., 'RGB1-U8' at 58.3% and 'RGB1-T24' at 60.1%, versus 72.5% for 'RGB1-DIN-U8' and 73.0% for 'RGB1-DIN-T24'). This demonstrates that the addition of DIN, UV, and particularly thermal channels substantially compensates for the loss of information in the underexposed RGB stream, even if it does not fully close the gap to ideal lighting.

- **Overexposed RGB ('RGB5'):** The system also maintained a substantial degree of robustness to overexposure, with the 'RGB5-DIN-T24-U8' configuration achieving a mean IoU of 84.2%. Although this is lower than the ideal-light result and marginally lower than the corresponding underexposed configuration (84.9%), it nonetheless represents a significant improvement over setups with less modalities in overexposed conditions. This indicates that, while the combination of DIN, UV, and thermal channels can substantially offset the loss of

information in overexposed RGB, some performance gap persists due to the challenges of information loss from saturation.

4.6.4 Challenging Classes: Contribution of Thermal and UV Streams

- **Partially decayed 'bad' fruit:** The two auxiliary channels, thermal (T24) and near-UV (U8), contribute in complementary ways. Thermal imagery is highly informative for 'bad' classes because incipient decay alters a fruit's metabolic heat and surface evaporation, producing local temperature contrasts. Conversely, near-UV sensing is sensitive to surface chemistry, revealing how different materials reflect or fluoresce under UV light. For instance, for 'Lemon Bad', the baseline RGB3-DIN model achieves an IoU of 47.1%. Adding the thermal channel (RGB3-DIN-T24) boosts this score significantly to 72.1%, and adding the UV channel (RGB3-DIN-U8) also provides a substantial improvement to 70.2

- **Plastic 'fake' fruit:** The auxiliary channels are also effective at identifying plastic fruit, which has a distinctive radiometric signature. For a particularly challenging class like 'Lemon Fake', the baseline RGB3-DIN model struggles at 29.6% IoU. Adding the thermal channel (RGB3-DIN-T24) is highly effective, causing performance to jump to 88.2%. While thermal is broadly useful, the UV channel provides a distinct advantage for specific categories, most notably 'Apple Green Fake', where the U8 stream (94.1%) outperforms the T24 stream (90.6%) under ideal RGB3 lighting. This suggests that UV cues are particularly effective for classes characterised by artificial surface properties.

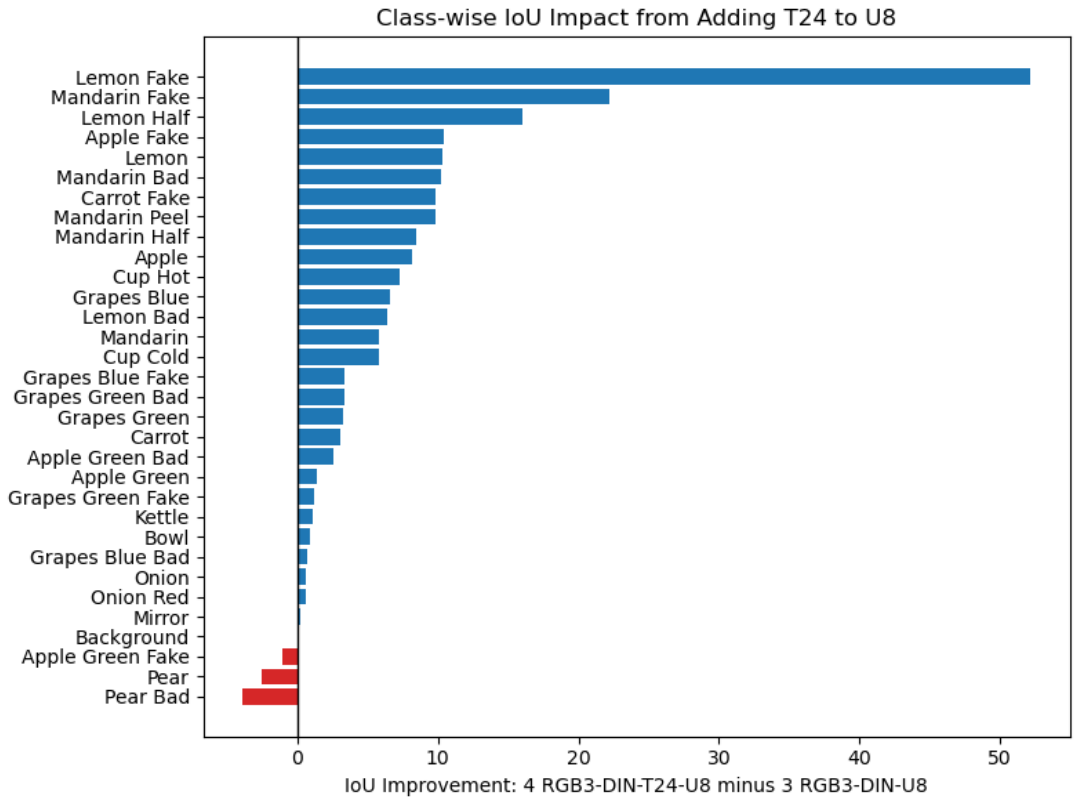
- **Fused multi-stream performance:** When all cues are provided (RGB-DIN-T24-U8), the network generally exploits the most salient stream per class. This configuration yields the highest overall mean IoU of 88.3% (with RGB3), surpassing both the T24-only (86.6%) and U8-only (81.9%) three-stream models. This fusion enables further gains in some cases; for instance, the IoU for 'Grapes Blue Fake' rises to 95.9% in the four-stream setting, surpassing both T24-only (93.9%) and U8-only (92.5%) results. However, for classes where one auxiliary stream is overwhelmingly dominant, adding the second can dilute the signal; the IoU for 'Apple Green Bad', for example, is higher with T24 alone (92.0%) than in the four-stream model (75.3%). Nonetheless, the aggregate metrics confirm that the four-stream model provides the most robust and balanced overall performance. An overview of the class-wise impact of adding UV to RGB3-DIN-T24 and T24 to RGB3-DIN-U8 is shown in Figure 4.10. While the addition of thermal data marginally impacts three classes negatively, the addition of UV has a more severe negative impact on specific classes.

4.6.5 Computational Requirements and Throughput

In addition to achieving high segmentation accuracy, the proposed architecture also enables real-time inference speeds. On a single RTX 3090 GPU, the four-modality configuration runs at 55 frames per second at a resolution of 640×480 pixels. The parameter count ranges from 11 million (for two-modality input) to 24 million (for the most comprehensive four-stream input), and computational cost scales from 10.8 GFLOPs to 17.3 GFLOPs. Even the most demanding setting maintains high throughput and can process full-resolution frames without significant latency, meeting the requirements of online robotic inspection and sorting systems. The mean rank metric, which summarises overall relative performance across all evaluated configurations, reaches a best value of 4.0 for the top fusion model, further underscoring the method's competitive standing.

4.6.6 Comparative Analysis of Fusion Strategies

To systematically evaluate our proposed sigmoid gating approach against state-of-the-art transformer-based attention mechanisms and assess multimodal integration strategies on the MM5 dataset, we conducted a comprehensive comparison of fusion architectures. Our primary goal was to directly contrast our lightweight per-pixel sigmoid gating with the cross-attention Feature Fusion Module (FFM) from the CMX framework—a representative transformer-based attention mechanism—thereby highlighting the trade-offs between accuracy, computational efficiency, and inference speed. Specifically, we aimed to demonstrate (a) the distinction between fusing data at the input level versus fusing features later in the network, (b) the performance and efficiency advantages of our sigmoid-based gating compared to

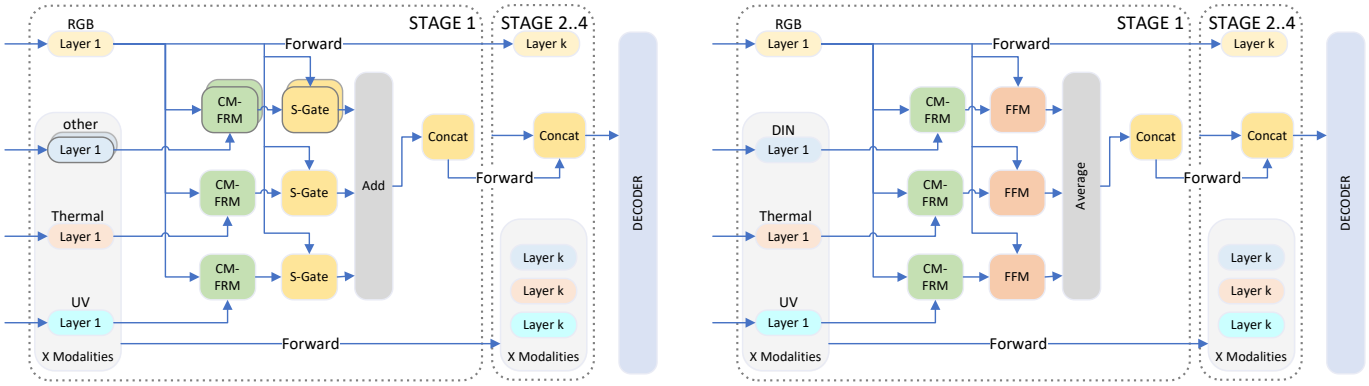


(a) Adding T24 to U8 fusion

(b) Adding U8 to T24 fusion

Figure 4.10: Class-wise IoU impact of adding T24 (a) or U8 (b) to multimodal fusion.

the more computationally intensive attention-based fusion employed in FFM, and (c) the relative merits of stage-wise intensity fusion versus pure feature-level fusion. This comparison directly addresses the question of whether simpler gating mechanisms can match or exceed the performance of complex transformer attention while maintaining real-time capability. Specifically, we compared the following approaches:



(a) GF-Net Gated

(b) CMX FRM/FFM - PAV

Figure 4.11: Architectural comparison of multimodal fusion strategies. (a) Our proposed GF-Net Gated model without SWIF, which fuses an auxiliary modality (X) with the RGB stream using a sigmoid gate. (b) A state-of-the-art baseline, CMX FRM/FFM - PAV, which employs a transformer-based cross-attention Feature Fusion Module (FFM).

GF-Net SWIF-Gated (DIN): Depth-intensity (DIN) features are merged using the stage-wise fusion (SWIF) module as shown in Figure 4.4, while all other modalities are fused via a learnt sigmoid gate as shown in Figure 4.2.
GF-Net Gated (D_FocusN+I): A per-pixel sigmoid gate is applied to the separate depth and intensity streams, as well as all other modalities, as shown in Figure 4.11a.
GF-Net Gated (DIN): A per-pixel sigmoid gate is applied to the fused DIN stream and all other modalities.
CMX FRM/FFM - PAV: We apply the cross-attention Feature Fusion Module (FFM) to each extra modality alongside RGB, then average the resulting feature maps before concatenation. In our comparisons, the 'parallel average' (PAV) strategy outperformed sequential, summation, concatenation, hierarchical, and simple gating variants, offering the best balance of accuracy and efficiency without overly complicating the architecture. The CMX FRM/FFM - PAV configuration represents a state-of-the-art transformer-based attention fusion approach, utilising the cross-attention mechanisms from CMX for each auxiliary modality. This serves as our primary baseline for evaluating whether our proposed lightweight sigmoid gating can achieve comparable or superior performance while reducing computational overhead. An overview of the architecture is shown in Figure 4.11b.
The quantitative results for these configurations are presented in Table 4.4, and a detailed class-level comparison is provided in Table 4.8.

Comparison Results

The gated stage-wise intensity fusion (SWIF) method consistently yielded the highest overall mean IoU, achieving 88.3% and rank 3.6 under ideal lighting ('RGB3-DIN-T24-U8'), and maintaining strong performance across adverse lighting scenarios (e.g., 84.9% for underexposed RGB and 84.2% for overexposed RGB). This approach also achieves these results with a reduced computational cost (17.3 GFLOPs) compared to the FFM/PAV (attention-based fusion) baseline, which reached 87.8% mean IoU at a higher cost (20.95 GFLOPs). Performance differences are particularly pronounced for under-represented or visually ambiguous categories, where explicit spatial gating and geometric cue enhancement enable more robust separation. The best results for underexposed RGB, achieving a mean IoU of 85.1% and rank 7.9, and for overexposed RGB, achieving a mean IoU of 86.1% and rank 5.7, were achieved by the network using only sigmoid gating and data-level fusion, highlighting the advantages of sigmoid gating in multimodal feature fusion as well as data-level fusion.
Notably, per-pixel gating with separate depth and intensity streams (D_FocusN+I) produced the highest class-wise IoU for certain categories. For example, segmentation performance for 'Apple Green Bad' improved dramatically from 75.3% (DIN-based fusion) to 89.2% with the D_FocusN+I variant, a gain of over 13.9 percentage points. In this variant, ADMRE-processed [1] depth (with normals) and NIR intensity are processed as separate streams before

Table 4.4: Class-wise segmentation results for representative fusion architectures using the MiT-B0 backbone and each network trained on 500 epochs. Each column group corresponds to a different fusion strategy: **GF-Net SWIF-Gated** (stage-wise intensity fusion with per-pixel gating), **GF-Net Gated** (per-pixel gating on fused DIN or on separate D_FocusN+I streams), **CMX FRM/FFM - PAV** (feature-rectify and channel-wise fusion with parallel average combination), and a downsampled variant (**GF-Net SWIF-Gated**, DIN at 320×240 resolution). Results are reported under three lighting conditions (underexposed 'RGB1', ideal 'RGB3', overexposed 'RGB5'). All values are the mean IoU per class. The bottom rows report the mean rank for each method, with lower values indicating stronger and more consistent performance across classes, as well as the average scores. This table substantiates the observed advantages of stage-wise, per-pixel gated fusion for robust multimodal segmentation, especially in adverse lighting and quantifies the trade-offs in accuracy, computational complexity, and efficiency among the variations.

	GF-Net SWIF-Gated (DIN)			**GF-Net Gated (D_FocusN+I)**			**GF-Net Gated (DIN)**			**CMX FRM/FFM - PAV (DIN)**			**GF-Net SWIF-Gated (DIN - 320x240)**		
Class	4 RGB1-DIN-T24-U8	4 RGB3-DIN-T24-U8	4 RGB5-DIN-T24-U8	RGB1-IAIP-D_FocusN-T24-U8	RGB3-IAIP-D_FocusN-T24-U8	RGB5-IAIP-D_FocusN-T24-U8	RGB1-DIN-T24-U8	RGB3-DIN-T24-U8	RGB5-DIN-T24-U8	RGB1-DIN-T24-U8	RGB3-DIN-T24-U8	RGB5-DIN-T24-U8	RGB1-DIN-T24-U8	RGB3-DIN-T24-U8	RGB5-DIN-T24-U8
Mean IoU	84.9	88.3	84.2	82.3	**88.4**	81.7	85.1	85.1	86.1	83.0	87.8	84.5	76.5	82.5	78.7
Freq IoU	99.4	**99.6**	99.4	99.3	**99.6**	99.4	99.4	99.5	99.5	99.4	99.5	99.4	99.1	99.2	99.1
Mean Pixel Acc	92.4	**93.9**	91.1	90.1	93.5	88.5	92.4	91.2	92.0	90.5	93.4	90.8	85.0	89.1	86.3
Pixel Acc	99.7	**99.8**	99.7	99.6	**99.8**	99.7	99.7	99.7	99.7	99.7	**99.8**	99.7	99.5	99.6	99.5
Mean Rank	8.4	**3.6**	7.5	10.1	4.0	9.2	7.9	6.4	5.7	9.4	4.0	7.1	13.5	10.4	12.8
FPS	55	55	55	41	41	41	52	52	52	37	37	37	91	91	91
Parameters	24M	24M	24M	29M	29M	29M	23M	23M	23M	23M	23M	23M	24M	24M	24M
Gflops	17.3	17.3	17.3	19.27	19.27	19.27	16.55	16.55	16.55	20.95	20.95	20.95	4.4	4.4	4.4

gating, rather than being fused at the data level. This substantial improvement can be attributed to the preservation and independent utilisation of geometric, intensity, and unaltered RGB data, which likely capture distinct cues not adequately represented when modalities are merged early. While there are some improvements for particular classes, especially under good lighting conditions, the overall performance remains similar and underperforms when light conditions are not ideal. Thus, this approach increases architectural complexity and does not improve network performance over the data-level fusion, further underpinned by the results of the network using only sigmoid gating and data-level fusion.

Importantly, the SWIF-Gated model sustains real-time throughput, achieving 55 fps at 640×480 pixels in the most demanding four-modality configuration, while maintaining lower GFLOPs than both the FFM/PAV approach and the five-stream gated variant. The five-stream gated model, incorporating five independent gated streams, operates at 41 fps, demonstrating the trade-off between modality count and inference speed. Additionally, the downsampled version at 320×240 resolution further reduces computational cost to 4.4 GFLOPs but incurs a substantial accuracy loss of 5.8% in mean IoU, highlighting the balance between efficiency and segmentation quality. Across all lighting conditions, all fusion approaches exhibited strong resilience to both underexposure ('RGB1') and overexposure ('RGB5'). However, stage-wise and per-pixel gated models consistently maintained higher accuracy on critical classes. They achieved superior overall mean IoU, while incurring minimal computational overhead compared to more complex attention-based fusion modules. These results confirm that stage-wise, per-pixel gated fusion offers more effective integration of auxiliary modalities than channel-wise or multi-stream attention mechanisms, particularly under challenging imaging conditions. Furthermore, by quantifying the trade-offs between segmentation accuracy, inference speed, and model complexity, our findings support the use of lightweight gating as a scalable solution for real-time multimodal semantic segmentation. A scatter plot illustrating the relationship between mean IoU and inference speed for the compared networks is presented in Figure 4.12.

These findings establish a rigorous baseline for future multimodal fusion architectures on MM5 and validate the

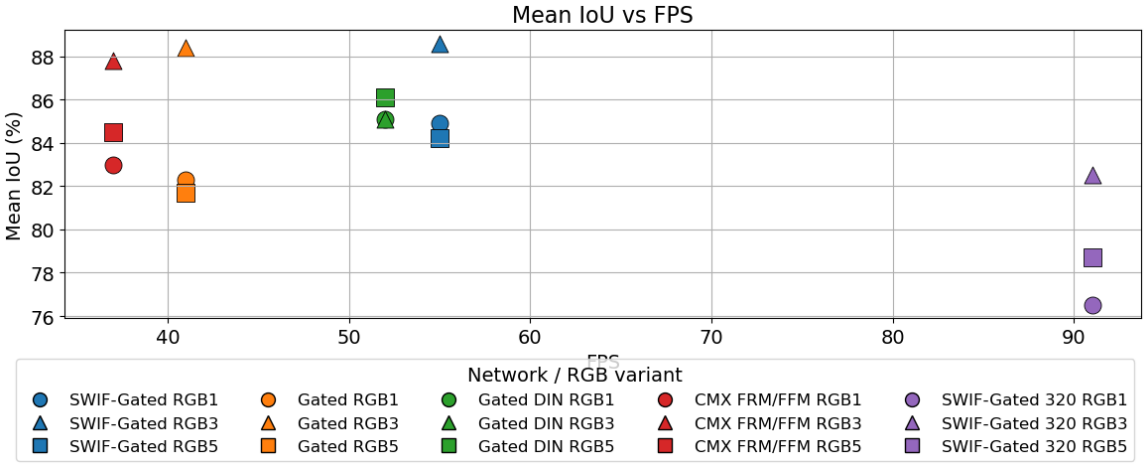


Figure 4.12: Mean IoU vs FPS scatter plot of the compared networks with all modalities as shown in Table 4.4. RGB1 being the underexposed, RGB5 the overexposed and RGB3 the ideal lighting.

effectiveness of content-adaptive, pixel-wise gating as a robust, efficient alternative to traditional attention-based fusion. However, as detailed in Section 4.7, the learnt gates specialise to training conditions and lack runtime adaptability when modalities are unexpectedly removed, highlighting an important limitation for future work to address.

4.7 Failure Case Analysis and Modality Importance

This section provides a rigorous account of where the proposed two-gate fusion system succeeds and where it fails, distinguishing persistent limits under full sensing from catastrophic collapses under modality ablation. We quantify effects across 12 evaluation scenarios and 32 semantic classes, yielding 384 class-scenario assessments, and we cross-reference these with class-level vulnerability profiles.

4.7.1 Failure Case Analysis with Full Multimodal Input

Evaluation across 76 test scenes per lighting condition (228 scenes total) reveals that despite achieving 99.72% pixel-level accuracy, systematic performance variations emerge across semantic categories. The overall error rate of 0.278% comprises boundary localisation errors (0.156%) and misclassifications (0.121%), with performance ranging from 99.78% under ideal illumination (RGB3) to 99.68% under challenging conditions (RGB1). The high accuracy on static background regions (99.87% across 228 instances) demonstrates effective object-background separation, yet specific object categories exhibit persistent failures. Analysis of 759 object instances across 32 semantic categories identifies that 16 classes exceed 3% misclassification rates. Degraded produce shows severe challenges: Pear Bad (23.55% error under RGB3; 12 instances; 28.55% overall), Mandarin Peel (8.72% under RGB3; 9 instances; 31.42% overall), and Apple Green Bad (13.68% under RGB3; 15 instances; 20.21% overall). These failures concentrate in regions with physical ambiguities that challenge multimodal sensing: severe occlusions where overlapping objects create ambiguous depth boundaries; specular reflections from Mirror (0.079% error, RGB3) and metallic Kettle (0.28% error, RGB3) that corrupt thermal and UV readings in adjacent regions; and gradual state transitions in degrading organic material where no discrete boundary exists. Thermal complexities compound these challenges—Cup Hot (1.50% error, RGB3) and Cup Cold (1.18% error, RGB3) show localised failures where heat radiation affects neighbouring objects' thermal signatures, creating phantom temperature readings that propagate classification errors. Cross-lighting analysis reveals substantial performance variance: Mandarin Peel exhibits 46.42 percentage point variation between conditions (RGB5: 55.14% vs. RGB3: 8.72%, while Apple Green Bad shows 24.87 points (RGB1: 35.91% vs. RGB5: 11.05%), indicating that certain failure modes are strongly illumination-dependent despite five-modality fusion. These empirical findings establish that 3.1% of semantic categories (1 of 32; Pear Bad) consistently fail to achieve 80% accuracy across all lighting conditions, while 25% (8 of 32) fall below 80% in at least one condition, delineating performance boundaries where physical ambiguities exceed the multimodal fusion

Table 4.5: Classes with persistent high misclassification rates under full multimodal fusion. Values show pixel-level accuracy and misclassification rates across lighting conditions.

Class	RGB1 Acc.	RGB3 Acc.	RGB5 Acc.	Avg. Error
Pear Bad	60.9%	76.5%	77.0%	28.5%
Mandarin Peel	69.6%	91.3%	44.9%	31.4%
Apple Green Bad	64.1%	86.3%	89.0%	20.2%
Lemon Bad	82.5%	87.4%	77.1%	17.7%
Mandarin Half	80.8%	84.7%	74.0%	20.2%
Mandarin Bad	79.8%	88.1%	81.2%	17.0%
Lemon Half	87.7%	76.0%	78.2%	19.5%

capabilities.

Class-Specific Performance Analysis

Analysis of 759 object instances across the 228 test scenes reveals persistent failure patterns for specific semantic categories. Table 4.5 presents the classes with consistent misclassification rates exceeding 10% with full multimodal fusion.
The degraded produce categories exhibit average error rates 5-20 times higher than their fresh counterparts, a disparity attributable to the subtle sensory cues required for decay detection—minute temperature variations and early-stage visual degradation that manifests as slight discolouration or texture softening that is barely distinguishable. Mandarin Peel demonstrates extreme performance instability (44.9% to 91.3% accuracy across lighting conditions). However, this variance partially reflects training artefacts from limited representation—only three evaluation instances across 76 test scenes—making the model sensitive to individual scene variations rather than learning robust class features. The underrepresentation is systemic across challenging categories: Mandarin Bad (5 evaluation instances), Mandarin Half (3 instances), and Pear Bad (4 instances) all exhibit high variance, correlating inversely with their training exposure. The performance failures concentrate in categories where either the distinguishing features approach sensor noise floors or training data inadequacy prevents robust feature learning.

Confusion Pattern Analysis

Systematic analysis of pixel-level predictions reveals two distinct error types: boundary errors occurring within 1-3 pixels of object edges and true misclassifications beyond this boundary zone. The network achieves 99.78% accuracy under optimal conditions (RGB3), with RGB1 at 99.68% and RGB5 at 99.71%. Analysis of the RGB3 configuration shows boundary errors account for approximately 0.137% of predictions while true misclassifications represent 0.087%, demonstrating that most errors occur at object boundaries rather than from semantic confusion. Figure 4.13 visualises the true misclassification patterns under optimal RGB3 conditions, excluding boundary errors. The analysis of the misclassified pixels across all lighting conditions reveals three dominant failure modes:

- **State-based confusion (≈38% of misclassifications):** Fresh-to-degraded transitions dominate semantic errors, with *Lemon*↔*Lemon Bad*, *Pear*↔*Pear Bad*, and *Mandarin*↔*Mandarin Bad* collectively accounting for 31,883 misclassified pixels. The confusion shows strong lighting dependence: RGB1 produces ≈34% more state-based errors than RGB3, indicating that underexposure specifically compromises decay signature detection despite thermal and UV modalities.
- **Background–object confusion (≈18% of all errors, Mirror/Bowl/Kettle):** These reflect pixels well within object interiors that are predicted as background. Across *Mirror*, *Bowl*, and *Kettle* we observe 35,530 object→background errors in total (Mirror 23,298; Bowl 7,917; Kettle 4,315). Including the reverse background→object direction brings this triad to 38,344 pixels. The confusion varies with illumination: *Mirror*→*Background* increases by 74% from RGB3 to RGB5 (5,640→9,831), consistent with specular-geometry effects.

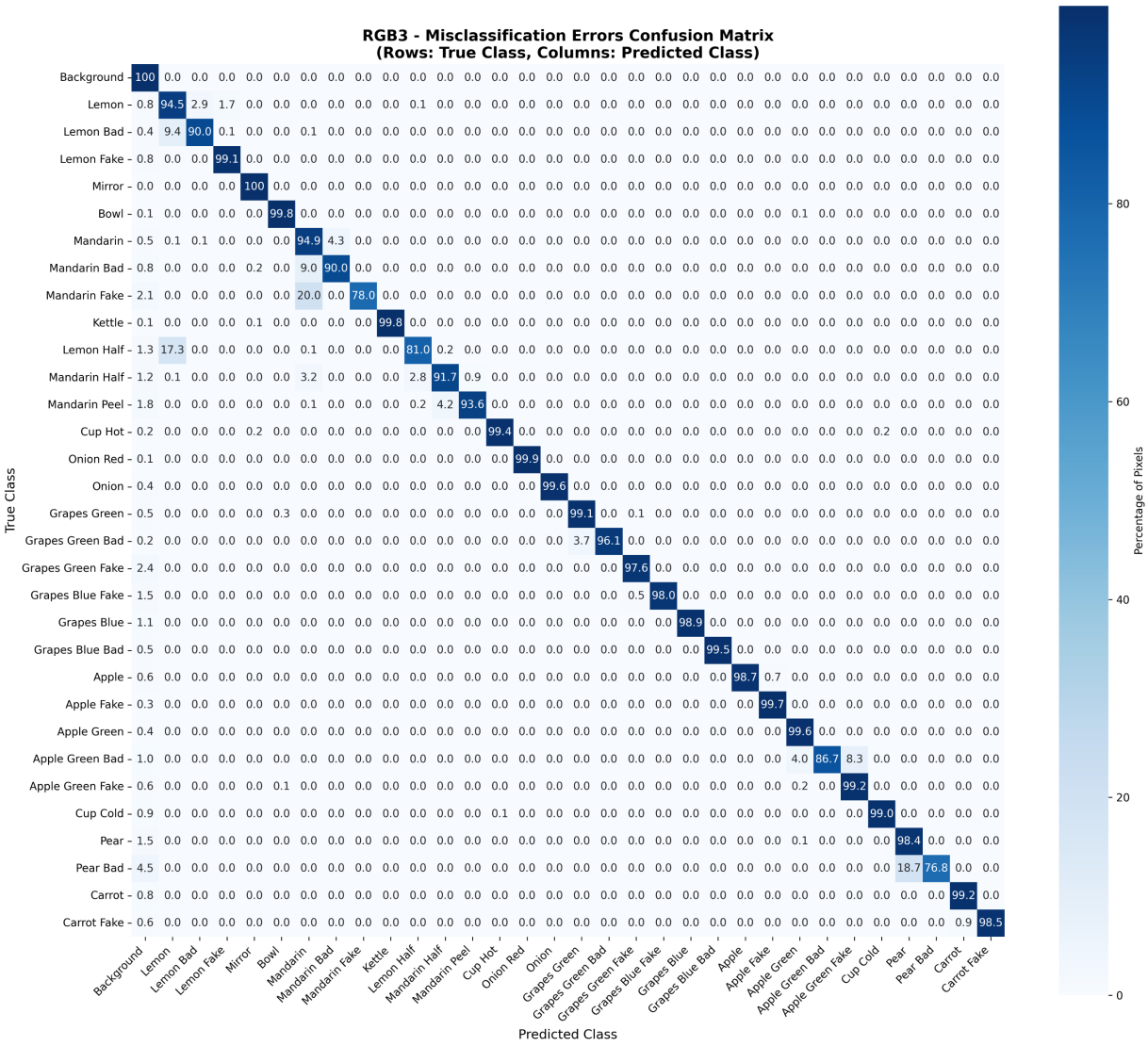


Figure 4.13: Confusion matrix under optimal lighting (RGB3) showing systematic misclassification patterns between class pairs. Boundary errors within 3 pixels of edges are excluded to focus on semantic confusion rather than localisation errors.

- **Material mimicry (≈7.7% of all errors):** Authentic-to-synthetic confusions collectively account for 15,054 pixels. Confusions show a directional bias toward synthetic→real rather than real→synthetic. For example, *Apple Green Fake→Apple Green Bad* accounts for 5,569 pixels (RGB1: 3,269; RGB3: 1,334; RGB5: 966). Aggregated over all "Fake" pairs, synthetic→real totals 9,080 pixels vs. real→synthetic 5,974. This suggests that certain artificial materials produce signatures (e.g., IR emissivity/UV response patterns) that more closely resemble degraded organic states, leading the model to default to "real/degraded" when cues are ambiguous.

Cross-lighting stability analysis reveals significant variation in confusion patterns across lighting conditions. The Mirror→Background confusion shows the highest variance (RGB1: 7,827 pixels; RGB3: 5,640; RGB5: 9,831), a $+74.3\%$ increase from RGB3 to RGB5 (half-range $\pm 2,096$ px). Other background-related pairs vary less, with half-range values of ± 188 px for Background→Mirror (731/355/550) and ± 282 px for Background→Carrot (800/236/453).

Spatial Error Distribution

Pixel-level error localisation reveals that failures concentrate in predictable spatial regions rather than a random distribution. Figure 4.14 illustrates the spatial patterns of boundary errors versus true misclassifications for a representative class.

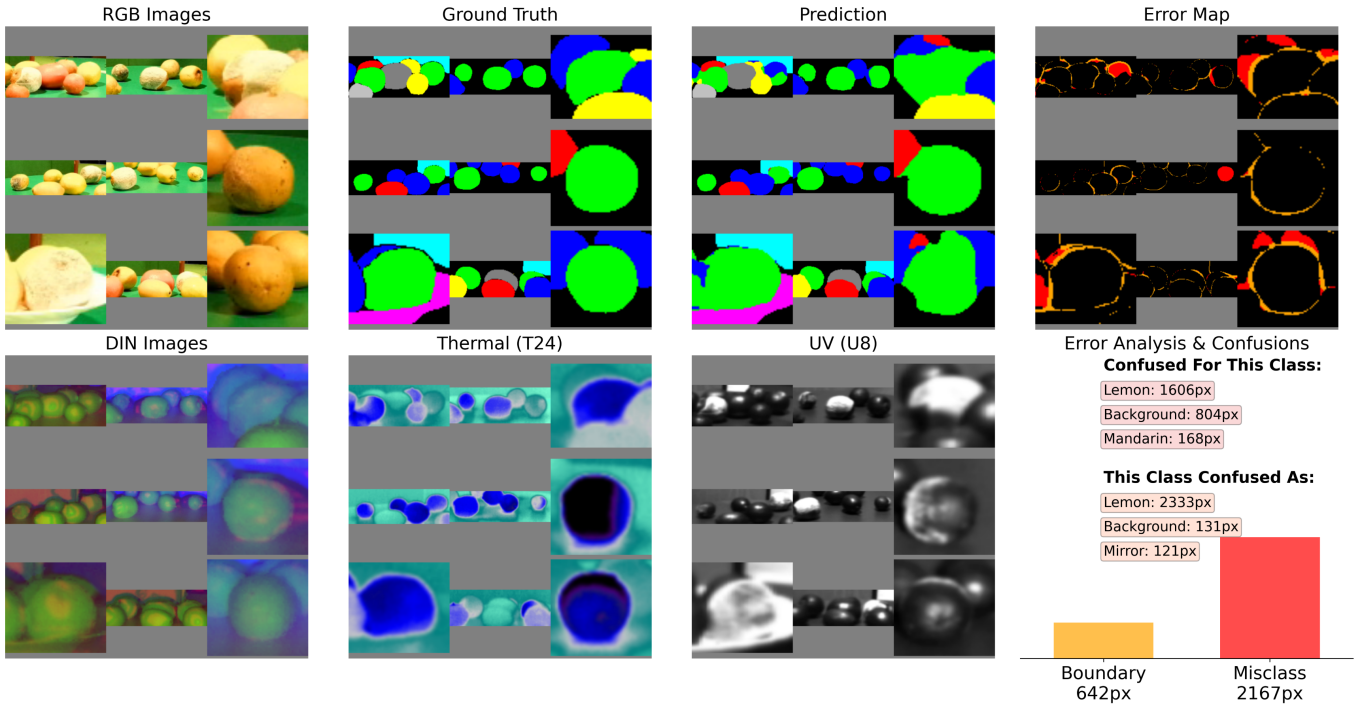


Figure 4.14: Spatial error distribution for the class Lemon Bad under RGB3 conditions. The top row shows RGB input, ground truth, prediction, and error map (orange: boundary errors within 3 pixels of edges; red: misclassifications). The bottom row shows the corresponding DIN, thermal, and UV modalities. Error concentration at object boundaries and decay transitions is evident, with 642 boundary pixels versus 2,167 misclassification pixels. GT: Lemon (Blue), Lemon Bad (Lime), Lemon Fake (Red), Mirror (Cyan), Bowl (Magenta), Mandarin (Yellow), Mandarin Bad (Light Grey), Mandarin Fake (Grey).

Error concentration analysis across all 32 classes and the three RGB setting reveals four primary failure regions:

- Object boundaries and annotation artefacts (56% of all errors): Boundary-error pixels within 3 px of edges total 109,515. The confusion matrix shows a strong asymmetry in mirror regions: Mirror→Background is 23,298 pixels, whereas the reverse Background→Mirror is 1,636. Similar asymmetries appear for bowls and kettles: Bowl→Background 7,917 vs. Background→Bowl 1,041, and Kettle→Background 4,315 vs. Background→Kettle 137. These patterns are consistent with reflective and intricate boundaries where annotation fragmentation and local appearance cues can diverge from the network's spatial coherence. Given Mirror's 99.92% accuracy under optimal lighting (RGB3), many such pixels likely reflect annotation/edge effects rather than substantive detection failures.

- Decay transition zones (21% of all errors; 47% of misclassifications): Across all Good/Bad pairs, state-based confusions sum to 39,944 pixels. For the highlighted pairs: Lemon Bad↔Lemon 13,895, Pear Bad↔Pear 9,051, and Mandarin↔Mandarin Bad 8,937. As shown in Figure 4.14, these errors are scattered along gradual decay gradients where no discrete boundary separates states.

- Background-to-object confusions (12% of all errors): Background→Object totals 23,422 pixels overall, with notable contributors including Background→Mirror 1,636, Background→Bowl 1,041, and Background→Kettle 137. These arise where extreme intensities or ambiguous boundaries yield object-like cues in the background.

- Inter-class confusions (11% of all errors): The remaining errors (beyond boundary effects, state-based pairs, and Background→object) occur at contact zones between spectrally similar materials where thermal cues blend and depth discontinuities weaken.

4.7.2 Modality ablation study

To quantify the contribution of each modality and diagnose robustness under sensor loss and degradation, we conducted comprehensive ablations across three illumination settings: RGB1, RGB3, and RGB5, corresponding to underexposed, optimal, and overexposed capture conditions, respectively. We evaluated both complete modality removal and controlled corruptions that emulate realistic sensor failures.
We evaluated 21 scenarios: one baseline (Full), four drop ablations (Drop_DIN, Drop_T24, Drop_U8, Drop_RGB), and sixteen noise ablations (four per modality). Noise ablations comprise one basic corruption and three advanced types. The basic `Noise` applies lightweight additive Gaussian perturbations with modality-specific but globally fixed scaling, followed by clamping to native data ranges. The advanced types— `Gaussian`, Salt & Pepper, and `Speckle`— implement adaptive Gaussian noise (per-channel scaling based on channel statistics with modality-specific minimum thresholds), impulse salt-and-pepper noise (modality-specific corruption probabilities), and multiplicative speckle noise (modality-specific intensities) respectively, with appropriate range clamping. This yields 21 total scenarios: 1 baseline + 4 drop + 4 basic + 12 advanced. Implementation specifics can be found in 4.B.2. Figure 4.15 illustrates the corruption types applied to the RGB3 modality.

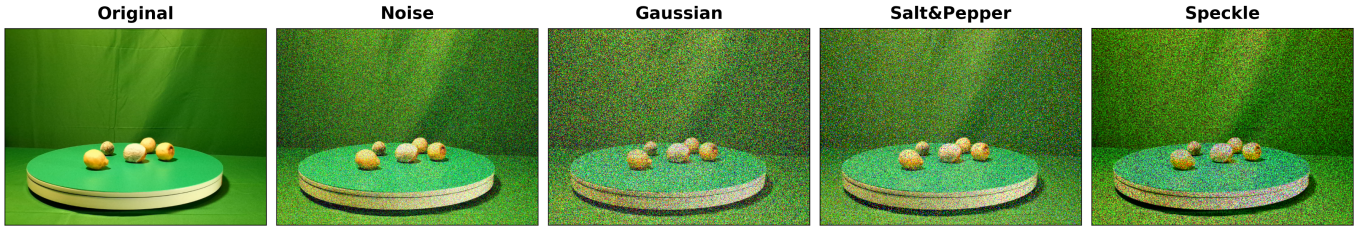


Figure 4.15: Examples of the implemented noise types. Applied to RGB3; columns show Original, Noise, Gaussian, Salt & Pepper, and Speckle.

Quantitative impact of modality removal

Across lighting settings, removal of the RGB stream yields the largest average degradation, followed by thermal, DIN, and UV, matching the ranking by average mIoU loss. For the four complete drops, the mean degradations are 83.4% (Drop_RGB), 63.3% (Drop_T24), 56.5% (Drop_DIN), and 30.1% (Drop_U8), with the single most severe configuration-specific collapse observed for RGB3_Drop_RGB at 90.2% degradation.

Sensor degradation scenarios and noise robustness

The noise corruption experiments reveal that degradation severity closely mirrors the dropout hierarchy. Most critically, thermal speckle noise causes 57.7% mIoU degradation—approaching the 63.3% degradation from complete thermal loss—indicating that corrupted thermal data can be nearly as detrimental as its absence. RGB exhibits a similar vulnerability, with Gaussian noise inducing 50.3% degradation compared to 83.4% for complete RGB removal. Salt-and-pepper corruption on thermal (55.3% degradation) further confirms the critical role of thermal.
DIN and UV show different patterns of noise resilience. UV corruptions cause minimal degradation (typically under 5%), likely reflecting the network's selective use of UV cues—UV dropout causes only 30.1% degradation compared to 56.5% for DIN. Notably, specific classes, such as Apple Green Bad, actually improve when UV is removed or when Gaussian noise is introduced to DIN, suggesting that these modalities can provide conflicting signals for specific categories. This selective modality usage demonstrates that the fusion mechanism learns task-specific dependencies, prioritising RGB and thermal for most classes whilst reserving UV for specialised discrimination tasks such as synthetic material detection. The consistent vulnerability hierarchy across both dropout and corruption tests confirms these learnt dependencies are systematic features of the trained model.
Per-class analysis confirms that losses concentrate on classes whose discriminative cues are tightly coupled to the ablated modality. Under Drop_RGB, Apple declines from an average IoU of 0.968 to 0.000, Grapes Blue from 0.957 to 0.045, and Mirror from 0.989 to 0.459, illustrating the dependence of chromatically distinctive and texture-rich categories on RGB cues. Under Drop_T24, thermally separable categories collapse, for example, Cup Cold from 0.961 to 0.000 and Cup Hot from 0.960 to 0.213, while Grapes Blue falls from 0.957 to 0.287. Under Drop_DIN,

Figure 4.16: Heatmap of class-wise IoU changes under drops and corruptions, red cells indicate low IoU scores and green indicate a good IoU score. Ablation Scenarios (RGB1/3/5 per scenario) D = Drop, N= Noise, G = Gaussian, SP = Salt & Pepper, S = Speckle.

geometrically intricate structures are most affected, for example, Grapes Blue from 0.957 to 0.503 and Bowl from 0.930 to 0.270. UV removal is most consequential for certain synthetic material categories, for example, Apple Green Fake, which decreased from 0.930 to 0.538, and Grapes Green Fake, which decreased from 0.938 to 0.738. Figure 4.16 presents a heatmap of the class-level IoU data, and Table 4.9 in 4.C presents network-level data, including per-stage gate activations.

Cross-RGB robustness

To assess lighting consistency, we calculated the coefficient of variation [30] across RGB1, RGB3, and RGB5 for each scenario. The results indicate an uneven distribution of robustness, with 13 scenarios classified as Low, five as Medium, and three as High. The Full baseline and the thermal-centred scenarios Drop_T24 and Speckle_T24 are among the most stable (High robustness). By contrast, scenarios dominated by DIN or UV under noise tend to be in the Low group. Drop_RGB is rated as Medium, reflecting the severe collapse under RGB3 that is partly offset by milder degradation in RGB1 and RGB5. Overall, these findings align with the baseline cross-lighting stability reported above.

Failure Mechanisms Under Ablation and Degradation

In our multi-scale architecture, Stages 1 to 4 proceed from the highest to the lowest spatial resolution, with feature map dimensions of H/4×W/4, H/8×W/8, H/16×W/16, and H/32×W/32, respectively, where H×W represents the input dimensions. Analysis of gate activation patterns reveals lighting-dependent modality utilisation. Under underexposed conditions (RGB1), thermal gates show moderate activation (59.4%) while UV remains relatively inactive (40.5%). Under optimal lighting (RGB3), thermal activation increases to 82.6% while UV is strongly suppressed (25.2%). Under overexposure (RGB5), thermal reaches near-complete activation (99.4%) while UV increases to 59.1%, suggesting UV provides complementary information primarily under challenging overexposed conditions.

The gating dynamics vary significantly across network stages. Early encoder stages (1-2) exhibit adaptive, continuous-valued gating that responds to input conditions—Stage 1 UV gates vary from 0.189 to 0.922 across scenarios while Stage 2 shows the widest dynamic range (thermal: 0.003-0.979, UV: 0.341-1.000). In contrast, later stages (3-4) display binary switching behaviour, with Stage 3 fixed near saturation or suppression, and Stage 4 showing discrete lighting-dependent configurations.
Within the encoder, Stage 1 maintains consistently high thermal activation across all conditions (>0.998), while UV activation at this stage varies with lighting (RGB1: 0.623, RGB3: 0.434, RGB5: 0.591). Stage 2 shows a different pattern, with UV dominating under underexposed conditions (RGB1: UV=0.993 vs thermal=0.377) but both modalities becoming highly active under overexposure (RGB5: thermal=0.977, UV=0.735). This complementary gating suggests the network learns to extract different features from each modality at different spatial resolutions.
These gating patterns reveal learnt but static modality dependencies that explain the differential impact of RGB removal. Under optimal lighting (RGB3), the network learns to extract highly detailed features from RGB, relegating auxiliary modalities to supplementary roles—the gates essentially specialise rather than adapt. This specialisation becomes catastrophic when RGB is removed (8.6% mIoU), as the pre-trained gates cannot dynamically adjust to redistribute processing to the available thermal and UV channels. Conversely, under challenging conditions, such as overexposure (RGB5), the network learns from the outset to rely more heavily on auxiliary modalities (thermal gates at 0.994), making it more resilient to RGB removal (13.0% mIoU). This suggests that the gates encode fixed strategies optimised for specific lighting conditions, rather than adaptive mechanisms that can respond to runtime modality availability.

4.7.3 Comparative Analysis of Failure Modes

Our analysis reveals two distinct failure regimes that emerge under different operational conditions. When all modalities function normally, errors concentrate at semantic boundaries and ambiguous regions—achieving 99.88% overall accuracy with only 0.12% error rate. These errors comprise boundary localisation issues (0.115%) and true misclassifications (0.0015%), primarily affecting degraded produce categories where decay transitions lack discrete boundaries.
Modality loss triggers catastrophic, systematic failures that dwarf baseline errors. The severity follows a clear hierarchy: RGB removal causes the most severe degradation (75.5% loss for RGB1, 90.2% for RGB3, 84.6% for RGB5), thermal removal substantially impacts performance (65.7% for RGB1, 58.0% for RGB3, 66.3% for RGB5), DIN removal shows moderate to severe effects (67.1% for RGB1, 33.3% for RGB3, 69.2% for RGB5), while UV removal has the smallest but still significant impact (31.9% for RGB1, 25.9% for RGB3, 32.4% for RGB5). These failures concentrate in modality-dependent classes—Cup Cold drops from 96.1% to 0.0% IoU without thermal, while Apple Green Fake falls from 93.0% to 53.8% without UV signatures.
The gating analysis reveals why RGB3 suffers most severely from RGB removal (90.2% degradation versus 75.5% for RGB1 and 84.6% for RGB5). Under optimal RGB3 conditions, the network develops specialised processing with high thermal reliance (gates: 0.826) while strongly suppressing UV (0.252), with RGB providing primary discriminative features. These learnt gate configurations remain fixed during inference; when RGB disappears, the pre-trained gates cannot dynamically redistribute the processing load. RGB5's near-complete thermal activation (gates: 0.994) combined with moderate UV activation (0.591) provides slightly better resilience, reducing RGB removal impact to 84.6%. RGB1, with moderate activation of both thermal (0.594) and UV (0.405), maintains the best resilience (75.5% degradation) due to its more balanced multi-modal processing strategy.

4.8 Conclusion

We have introduced GatedFusion-Net, a lightweight hierarchical fusion architecture that delivers state-of-the-art segmentation on the five-modality MM5 dataset at real-time speeds. By injecting a data-level Depth-Intensity-Normal (DIN) composite into the SegFormer backbone at every encoder scale, our model sharpens object boundaries and mitigates saturation or underexposure artefacts without extra memory overhead. Aligned thermal (T24) and ultraviolet (U8) streams are rectified via CMX's FRM and then gated per-pixel by learnt sigmoid masks, ensuring that only informative cues contribute to the final representation.
With 24M parameters and 17.3 GFLOPs, GatedFusion-Net achieves up to 74 fps (four-modality) and 55 fps (five-modality) on 640×480 inputs, while reaching a peak mIoU of 88.3% and 99.8% pixel accuracy. The network

maintains robust performance under under- and over-exposed RGB, where adding DIN raises mIoU from 60.1% to 85.6%, nearly matching the 86.6% obtained with ideal RGB, confirming that NIR and depth effectively compensate for degraded colour information. Thermal cues consistently yield the most significant standalone gains, especially for detecting rot, and UV aids in distinguishing synthetic replicas from genuine produce.

Our comprehensive failure analysis reveals important limitations alongside these achievements. Whilst the system maintains high accuracy under normal operation, certain semantic categories remain challenging—degraded produce classes such as Pear Bad exhibit error rates exceeding 23%, and complete modality loss triggers catastrophic failures with up to 90.2% performance degradation. The analysis of gate activation patterns indicates that our fusion learns static, lighting-specific strategies rather than adaptive mechanisms, explaining why RGB removal under optimal lighting causes more severe degradation (90.2%) than under challenging conditions (75.5% for RGB1). These findings demonstrate that, while our architecture achieves robust multimodal integration for real-world applications, the identified failure modes under sensor loss reveal opportunities for developing adaptive fusion mechanisms that can dynamically reconfigure when modalities become unavailable.

These findings confirm that stage-wise, transformer-based fusion can seamlessly integrate more than three modalities for both domestic and industrial inspection tasks, and that modality-wise gating provides a lightweight alternative to heavier attention modules. Future work will explore adaptive quality prediction to further down-weight unreliable streams and extend the framework to additional sensor types.

We acknowledge that our work is validated exclusively on the MM5 dataset, which comprises indoor scenes of produce (fruit and vegetables) captured under controlled lighting variations. The generalisation of our learnt weighting patterns to other multi-modal datasets, outdoor environments, or different object categories and sensor combinations remains to be explored empirically.

4.9 Future Work

While the proposed fusion architecture establishes a strong baseline for multimodal segmentation on the MM5 dataset, several directions remain for further exploration and improvement.

Addressing Class Imbalance: The MM5 dataset exhibits a long-tail distribution. Future work will focus on mitigating this imbalance through strategies such as stratified or synthetic data augmentation, class-balanced and focal loss functions, and few-shot adaptation techniques that could further improve segmentation performance for rare classes.

Advanced Regularisation and Training Protocols: Although our results indicate that smaller models (e.g., MiT-B0) benefit more from extended training, while larger backbones (e.g., MiT-B2) plateau more rapidly, exploring advanced regularisation strategies, including curriculum learning, more substantial data augmentation, or semi-supervised learning, could further improve generalisation and resource efficiency. Future experiments may also systematically compare the impact of early stopping, adaptive learning rates, and other optimisation techniques not applied in the present study.

Adaptive Fusion Mechanisms: Our failure analysis reveals that current gating learns static, lighting-dependent strategies rather than adapting to runtime sensor availability. Future architectures could incorporate three key improvements: structured modality dropout during training to encourage robust, redundant feature extraction across all sensors; explicit degradation detection mechanisms that dynamically adjust gating weights when input quality degrades; and regularisation techniques that enforce cross-modal redundancy for critical features, preventing catastrophic failure when individual modalities become unavailable.

Benchmarking and Transferability: As the MM5 dataset becomes a reference point for multimodal segmentation, future work will also focus on extending the dataset to include additional object classes, capturing the same classes in diverse environments, and acquiring sequences of video footage. These efforts will further expand the benchmarking capabilities of MM5 and enable a more comprehensive evaluation of model generalisability across varied conditions.

Code Availability

The code used in this paper will be made publicly available at
`https://github.com/martinbrennernz/MM5-Dataset`
upon publication of this work.

4.A Detailed Network Results

4.A.1 MiT-B0 500 Epochs

Table 4.6: Detailed IoU results for various modality combinations and lighting conditions trained for 500 epochs (best in each row bold) using a MiT-B0 backbone. DIN: Depth-Intensity-Normals fused; T24: processed thermal; U8: ultraviolet; RGB1: under-exposed RGB; RGB3: ideal RGB; RGB5: over-exposed RGB. "Bad" classes are partially rotten; "Fake" classes are replicas.

Class	2 RGB1-U8	2 RGB1-T24	2 RGB3-DIN	3 RGB1-DIN-U8	3 RGB3-DIN-U8	3 RGB5-DIN-U8	3 RGB1-DIN-T24	3 RGB3-DIN-T24	3 RGB5-DIN-T24	4 RGB1-DIN-T24-U8	4 RGB3-DIN-T24-U8	4 RGB5-DIN-T24-U8
Background	99.6	99.6	99.8	99.8	**99.9**	99.8	99.8	**99.9**	99.8	99.8	**99.9**	99.8
Lemon	56.0	60.0	64.4	70.5	70.8	60.1	73.7	79.2	69.9	79.3	**81.1**	71.9
Lemon Bad	44.1	36.9	47.1	66.6	70.2	52.1	64.0	72.1	62.9	**77.2**	76.6	68.1
Lemon Fake	19.9	31.9	29.6	6.2	35.1	18.8	84.5	**88.2**	78.5	87.2	87.3	78.5
Mirror	95.7	96.3	97.9	98.2	98.7	97.6	98.4	98.7	98.5	98.2	**98.9**	98.2
Bowl	86.4	86.7	90.1	91.5	92.1	90.5	91.0	91.7	90.3	92.0	**93.0**	92.5
Mandarin	74.5	78.6	83.2	72.3	78.6	67.9	83.4	82.0	78.4	83.1	**84.4**	71.8
Mandarin Bad	45.7	53.8	57.1	45.6	57.6	28.6	64.5	57.0	51.5	62.9	**67.8**	53.9
Mandarin Fake	81.5	86.6	88.0	54.4	65.8	57.0	**90.0**	84.7	89.1	86.1	88.0	62.2
Kettle	87.4	86.4	94.4	94.5	95.4	94.0	93.8	95.3	95.1	92.6	**96.5**	94.6
Lemon Half	29.9	25.3	33.8	40.3	57.6	49.7	77.1	71.4	**78.8**	64.9	73.6	71.4
Mandarin Half	29.3	28.8	**74.6**	35.5	63.5	49.8	68.9	67.5	60.5	65.0	72.0	66.1
Mandarin Peel	32.1	67.4	53.9	53.2	66.9	34.7	75.9	43.2	31.6	65.2	**76.7**	56.0
Cup Hot	43.8	44.4	81.0	92.0	88.7	89.1	94.1	95.2	93.8	93.8	**96.0**	93.9
Onion Red	76.6	78.1	95.7	82.6	95.9	62.9	89.8	**96.5**	94.7	92.6	**96.5**	94.7
Onion	82.5	83.0	96.0	91.1	96.5	95.6	90.2	96.7	95.8	94.0	**97.1**	95.9
Grapes Green	75.7	77.3	89.8	88.0	90.3	89.3	87.7	90.0	91.2	91.0	**93.6**	92.9
Grapes Green Bad	60.7	75.5	87.0	81.3	87.1	85.5	83.9	89.3	89.1	85.8	**90.5**	89.1
Grapes Green Fake	73.8	75.3	89.3	83.1	92.6	90.5	87.3	93.4	90.7	85.6	**93.8**	90.1
Grapes Blue Fake	82.8	85.8	93.8	74.7	92.5	89.6	88.6	93.9	95.5	88.2	**95.9**	93.2
Grapes Blue	10.2	11.3	93.5	70.2	89.2	84.6	90.3	93.8	95.6	94.1	95.8	**96.4**
Grapes Blue Bad	34.6	35.1	94.2	92.3	94.9	95.9	92.5	94.9	**96.2**	93.7	95.6	95.9
Apple	52.7	58.9	91.9	55.3	88.6	48.1	96.1	96.1	96.5	96.1	**96.8**	96.6
Apple Fake	64.7	59.4	89.5	58.4	85.2	66.9	95.3	93.8	94.6	95.0	**95.6**	94.6
Apple Green	67.9	64.2	89.5	74.8	85.8	80.1	94.4	**94.9**	94.0	82.4	87.2	83.7
Apple Green Bad	60.9	46.7	**94.8**	59.7	72.7	59.0	75.5	92.0	92.1	60.9	75.3	72.5
Apple Green Fake	66.3	70.5	83.4	93.3	94.1	94.0	81.8	90.6	92.7	85.4	93.0	**94.2**
Cup Cold	28.8	29.2	89.6	91.6	90.3	93.7	93.1	94.4	92.3	90.4	**96.1**	92.7
Pear	48.2	66.0	75.7	70.8	78.3	76.0	75.2	**79.6**	73.8	76.5	75.8	76.1
Pear Bad	43.9	57.2	75.7	68.5	77.8	69.9	75.9	**78.5**	71.8	74.3	73.9	76.5
Carrot	66.5	41.6	87.3	86.5	89.5	88.2	91.4	91.9	91.3	91.9	**92.6**	92.3
Carrot Fake	41.9	26.6	71.2	77.3	78.1	76.8	91.9	85.0	84.4	**92.6**	87.9	87.0
Mean IoU	58.3	60.1	80.7	72.5	81.9	73.0	85.6	86.6	84.7	84.9	**88.3**	84.2
Freq IoU	98.6	98.6	99.3	99.1	99.4	99.1	99.4	99.5	99.5	99.4	**99.6**	99.4
Mean Pixel Acc	71.1	72.4	88.7	82.5	88.9	82.5	92.3	92.3	91.3	92.4	**93.9**	91.1
Pixel Acc	99.1	99.1	99.6	99.5	99.7	99.5	99.7	**99.8**	99.7	99.7	**99.8**	99.7
Mean Rank	11.1	10.4	6.7	9.0	5.3	8.4	5.4	3.7	5.4	5.6	**2.2**	4.8
FPS	104	104	104	74	74	74	74	74	74	55	55	55
Parameters	11M	11M	11M	18M	18M	18M	18M	18M	18M	24M	24M	24M
GFLOPs	10.8	10.8	10.8	14.5	14.5	14.5	14.5	14.5	14.5	17.3	17.3	17.3

4.A.2 MiT-B2 250 Epochs

Table 4.7: Detailed IoU results for various modality combinations and lighting conditions trained for 250 epochs (best in each row bold) using a MiT-B2 backbone. DIN: Depth-Intensity-Normals fused; T24: Processed Thermal; U8: Ultraviolet; RGB1: Underexposed RGB; RGB3: Ideal RGB; RGB5: Overexposed RGB. Bad classes are partially rotten; Fake classes are replicas.

Class	2 RGB1-U8	2 RGB1-T24	2 RGB3-DIN	3 RGB1-DIN-U8	3 RGB3-DIN-U8	3 RGB5-DIN-U8	3 RGB1-DIN-T24	3 RGB3-DIN-T24	3 RGB5-DIN-T24	4 RGB1-DIN-T24-U8	4 RGB3-DIN-T24-U8	4 RGB5-DIN-T24-U8
Background	96.4	99.5	99.7	99.7	**99.8**	**99.8**	99.6	99.7	99.7	99.7	**99.8**	99.7
Lemon	50.1	54.9	63.3	68.7	68.1	63.9	72.7	68.7	71.3	75.9	**79.0**	69.4
Lemon Bad	32.6	50.1	53.6	71.8	66.7	60.3	65.7	57.6	66.8	**72.7**	71.1	68.2
Lemon Fake	21.9	61.5	39.3	41.4	41.4	38.5	86.5	74.6	80.4	86.0	**90.2**	81.7
Mirror	49.1	95.3	94.9	98.1	98.3	97.4	95.5	96.5	97.5	97.5	**98.9**	97.6
Bowl	68.3	77.5	89.5	89.3	91.6	89.0	84.4	86.7	87.2	90.3	**91.2**	86.1
Mandarin	70.0	64.7	84.7	75.7	80.6	72.6	82.1	80.6	76.4	**87.0**	85.7	73.5
Mandarin Bad	36.3	33.2	73.2	62.3	62.8	29.4	58.9	58.0	55.4	**75.7**	68.7	58.4
Mandarin Fake	59.5	57.0	83.1	62.8	75.8	79.5	82.7	75.7	66.7	88.1	**91.2**	84.9
Kettle	50.3	83.4	88.6	85.9	93.7	91.0	84.7	89.2	88.2	87.4	**94.3**	92.8
Lemon Half	39.3	68.3	51.8	42.1	56.7	32.0	69.1	51.1	64.1	57.8	61.4	**64.9**
Mandarin Half	56.3	17.0	69.4	33.5	64.3	45.2	77.0	67.7	57.4	57.0	**73.3**	65.0
Mandarin Peel	0.0	1.2	43.0	56.6	28.3	34.3	**71.5**	36.7	15.8	67.5	56.2	67.1
Cup Hot	34.9	86.0	65.2	71.9	85.6	86.9	90.8	90.1	93.2	93.2	**95.0**	93.9
Onion Red	68.3	71.6	94.6	78.9	94.2	79.7	88.2	92.7	93.4	85.3	**94.9**	93.3
Onion	77.7	89.0	94.6	88.3	95.5	94.7	89.0	93.9	94.4	82.6	**95.4**	93.7
Grapes Green	64.6	62.0	87.6	83.0	90.7	87.6	86.2	87.3	90.8	86.8	90.0	**91.9**
Grapes Green Bad	32.6	48.7	84.8	79.7	86.5	79.8	78.0	88.8	88.1	83.4	**90.3**	86.3
Grapes Green Fake	61.2	44.8	83.3	84.4	**92.0**	88.3	85.0	87.4	89.9	84.7	85.6	79.9
Grapes Blue Fake	54.1	29.8	79.5	88.5	79.0	93.3	89.5	93.6	93.9	89.7	86.2	**83.4**
Grapes Blue	23.5	35.1	59.0	71.7	48.4	95.0	89.9	92.1	94.0	88.1	**93.9**	93.3
Grapes Blue Bad	28.7	61.2	79.5	82.0	89.0	95.3	84.5	92.4	94.5	87.6	93.6	**94.9**
Apple	49.5	83.0	83.4	79.4	85.7	73.3	93.5	91.3	95.2	95.1	**95.8**	95.2
Apple Fake	45.7	69.1	82.2	74.6	82.2	81.7	92.5	88.4	93.8	93.8	**94.8**	94.7
Apple Green	54.6	61.1	83.6	71.5	84.8	84.8	**86.5**	82.8	85.7	86.1	84.7	82.4
Apple Green Bad	33.7	56.9	61.7	51.2	64.8	59.8	**82.0**	59.9	80.7	77.0	74.4	60.6
Apple Green Fake	44.6	56.3	91.3	91.5	93.6	93.7	91.9	80.5	89.1	89.7	**94.6**	93.1
Cup Cold	19.1	74.8	70.0	53.4	86.6	86.4	88.0	88.7	92.3	91.7	**94.7**	92.6
Pear	30.6	38.8	61.5	71.4	71.9	67.3	61.8	62.2	71.4	75.4	**78.9**	76.0
Pear Bad	13.0	30.2	59.4	70.6	78.1	77.9	51.4	53.5	69.4	74.5	**78.6**	78.4
Carrot	58.8	86.9	85.9	83.6	89.7	87.4	89.9	86.0	90.0	89.9	91.2	**91.6**
Carrot Fake	41.3	70.9	69.0	71.8	80.6	78.2	88.4	71.5	83.5	85.5	82.9	**89.7**
Mean IoU	45.8	58.2	75.3	73.0	78.3	75.7	82.4	78.9	81.6	83.8	**86.1**	83.6
Freq IoU	93.8	98.5	99.0	99.1	99.3	99.2	99.1	99.1	99.3	99.3	**99.5**	99.3
Mean Pixel Acc	63.9	69.7	85.5	83.7	86.9	85.7	91.8	87.4	90.3	92.4	**93.7**	91.9
Pixel Acc	95.8	99.1	99.4	99.5	99.6	99.5	99.5	99.5	99.6	99.6	**99.7**	99.6
Mean Rank	10.7	10.5	6.4	9.0	5.8	7.8	5.7	3.3	6.3	4.8	**2.9**	4.9
FPS	39	39	39	29	29	29	29	29	29	25	25	25
Parameters	67M	67M	67M	106M	106M	106M	106M	106M	106M	140M	140M	140M
GFLOPs	60.9	60.9	60.9	84.8	84.8	84.8	84.8	84.8	84.8	105.0	105.0	105.0

4.A.3 MiT-B0 Comparison

Table 4.8: Class-wise segmentation results for representative fusion architectures using the MiT-B0 backbone and each network trained on 500 epochs. Each column group corresponds to a different fusion strategy: **GF-Net SWIF-Gated** (stage-wise intensity fusion with per-pixel gating), **GF-Net Gated** (per-pixel gating on fused DIN or on separate D_FocusN+I streams), **CMX FRM/FFM - PAV** (feature-rectify and channel-wise fusion with parallel average combination), and a downsampled variant (**GF-Net SWIF-Gated**, DIN at 320×240 resolution). Results are reported under three lighting conditions (underexposed 'RGB1', ideal 'RGB3', overexposed 'RGB5'). All values are mean IoU per class. The bottom rows report the mean rank for each method, with lower values indicating stronger and more consistent performance across classes and the average scores. This table substantiates the observed advantages of stage-wise, per-pixel gated fusion for robust multimodal segmentation, especially in adverse lighting and quantifies the trade-offs in accuracy, computational complexity, and efficiency among the variations.

	GF-Net SWIF-Gated (DIN)			**GF-Net Gated (D_FocusN+I)**			**GF-Net Gated (DIN)**			**CMX FRM/FFM - PAV (DIN)**			**GF-Net SWIF-Gated (DIN - 320x240)**		
Class	4 RGB1-DIN-T24-U8	4 RGB3-DIN-T24-U8	4 RGB5-DIN-T24-U8	RGB1-IAIP-D_FocusN-T24-U8	RGB3-IAIP-D_FocusN-T24-U8	RGB5-IAIP-D_FocusN-T24-U8	RGB1-DIN-T24-U8	RGB3-DIN-T24-U8	RGB5-DIN-T24-U8	RGB1-DIN-T24-U8	RGB3-DIN-T24-U8	RGB5-DIN-T24-U8	RGB1-DIN-T24-U8	RGB3-DIN-T24-U8	RGB5-DIN-T24-U8
Background	**99.9**	**99.9**	99.8	99.8	**99.9**	99.8	99.8	**99.9**	99.8	99.8	**99.9**	99.8	99.7	99.7	99.7
Lemon	79.3	**81.1**	71.9	75.3	80.5	70.0	77.8	78.8	73.8	73.9	75.4	68.2	66.7	67.4	67.8
Lemon Bad	**77.2**	76.6	68.1	66.9	75.8	60.1	73.9	71.2	66.3	75.0	66.2	65.9	65.4	60.1	62.4
Lemon Fake	87.2	**87.3**	78.5	71.7	85.7	82.3	78.0	87.7	82.7	70.1	86.1	79.1	55.2	75.4	63.0
Mirror	98.2	**98.9**	98.2	98.0	98.6	98.0	98.3	98.6	98.3	98.3	98.8	98.1	97.0	95.9	96.1
Bowl	92.0	93.0	92.5	92.0	92.9	92.5	91.8	**93.4**	93.0	91.7	93.2	92.2	89.1	91.4	90.8
Mandarin	83.1	84.4	71.8	79.6	83.7	73.3	**86.6**	79.7	82.8	80.3	84.2	69.7	75.0	87.5	71.8
Mandarin Bad	62.9	67.8	53.9	60.8	63.1	34.8	**76.3**	56.4	56.6	57.7	65.7	46.9	58.4	77.8	56.3
Mandarin Fake	86.1	88.0	62.2	59.0	80.0	76.0	80.5	74.5	**90.3**	59.3	81.7	83.3	43.9	91.1	44.2
Kettle	92.6	**96.5**	94.6	90.8	95.2	94.0	92.2	95.5	95.2	93.0	95.5	95.3	91.0	95.1	94.6
Lemon Half	64.9	**73.6**	71.4	68.3	68.8	65.3	70.3	72.3	69.7	64.5	70.1	66.1	47.4	54.2	62.0
Mandarin Half	65.0	**72.0**	66.1	68.6	77.2	55.9	65.0	59.8	64.1	66.1	**77.3**	60.2	48.4	56.9	59.4
Mandarin Peel	65.2	**76.7**	56.0	62.5	80.7	41.7	65.9	56.6	54.5	58.8	**82.6**	47.5	35.2	36.2	25.8
Cup Hot	93.8	**96.0**	93.9	93.7	95.0	94.6	94.7	93.3	94.1	93.9	95.1	93.7	92.1	93.8	91.5
Onion Red	92.6	**96.5**	94.7	92.5	96.2	94.7	89.3	96.3	94.5	92.8	95.9	94.2	87.4	94.4	93.1
Onion	94.0	**97.1**	95.9	93.6	96.7	96.0	90.2	96.8	96.2	94.6	96.9	96.1	86.8	94.6	91.6
Grapes Green	91.0	93.6	92.9	88.6	92.8	92.1	83.3	92.8	93.3	87.0	92.7	92.3	88.0	89.3	86.0
Grapes Green Bad	85.8	**90.5**	89.1	85.6	87.5	85.2	85.3	90.2	89.7	86.5	90.9	88.2	86.1	87.2	86.1
Grapes Green Fake	85.6	**93.8**	90.1	72.7	93.6	88.5	69.4	86.8	93.0	80.5	94.0	92.2	85.2	86.3	80.1
Grapes Blue Fake	88.2	**95.9**	93.2	64.9	96.0	80.5	68.0	90.1	94.4	83.9	95.5	95.3	85.8	88.9	92.2
Grapes Blue	94.1	95.8	**96.4**	93.7	95.5	73.5	93.8	95.1	96.5	93.8	95.3	96.7	86.2	93.1	91.5
Grapes Blue Bad	93.7	95.6	95.9	93.2	95.1	96.1	93.9	95.6	**96.7**	94.1	95.8	96.6	92.5	94.6	93.9
Apple	96.1	96.8	96.6	95.6	96.1	96.5	96.0	96.7	**96.8**	95.4	96.9	96.5	92.6	94.6	93.8
Apple Fake	95.0	95.6	94.6	95.0	95.8	94.6	94.0	**96.1**	95.4	93.1	95.5	95.0	90.2	93.6	92.3
Apple Green	82.4	87.2	83.7	90.5	**92.6**	89.8	94.4	88.7	87.6	92.7	88.7	89.3	87.6	87.7	86.7
Apple Green Bad	60.9	75.3	72.5	85.2	**89.2**	82.5	92.1	77.8	77.3	77.9	79.9	79.0	56.2	61.6	66.6
Apple Green Fake	85.4	93.0	94.2	91.4	**94.5**	93.5	93.2	94.1	92.9	83.3	94.5	93.8	76.6	82.1	81.7
Cup Cold	90.4	**96.1**	92.7	92.4	95.5	94.2	93.2	94.3	94.9	93.8	94.9	94.7	90.7	93.6	90.3
Pear	76.5	75.8	76.1	68.0	77.1	76.2	76.1	68.2	**78.3**	73.1	72.6	78.9	62.0	72.7	69.3
Pear Bad	74.3	73.9	76.5	60.4	74.3	76.1	77.2	68.1	**77.8**	70.3	74.2	78.5	53.0	73.9	68.3
Carrot	91.9	92.6	92.3	91.6	93.2	89.2	92.3	91.8	92.6	90.9	92.9	92.2	88.9	88.9	87.9
Carrot Fake	**92.6**	87.9	87.0	90.9	90.0	77.8	91.4	84.9	87.9	89.1	89.2	88.4	86.7	80.8	83.2
Mean IoU	84.9	88.3	84.2	82.3	**88.4**	81.7	85.1	85.1	86.1	83.0	87.8	84.5	76.5	82.5	78.7
Freq IoU	99.4	**99.6**	99.4	99.3	**99.6**	99.4	99.4	99.5	99.5	99.4	99.5	99.4	99.1	99.2	99.1
Mean Pixel Acc	92.4	**93.9**	91.1	90.1	93.5	88.5	92.4	91.2	92.0	90.5	93.4	90.8	85.0	89.1	86.3
Pixel Acc	99.7	**99.8**	99.7	99.6	**99.8**	99.7	99.7	99.7	99.7	99.7	**99.8**	99.7	99.5	99.6	99.5
Mean Rank	8.4	**3.6**	7.5	10.1	4.0	9.2	7.9	6.4	5.7	9.4	4.0	7.1	13.5	10.4	12.8
FPS	55	55	55	41	41	41	52	52	52	37	37	37	91	91	91
Parameters	24M	24M	24M	29M	29M	29M	23M	23M	23M	23M	23M	23M	24M	24M	24M
Gflops	17.3	17.3	17.3	19.27	19.27	19.27	16.55	16.55	16.55	20.95	20.95	20.95	4.4	4.4	4.4

4.B Implementation Details

4.B.1 DIN Preprocessing Details

We generate the normal channel as follows:

1. Apply a bilateral filter to the depth focus image to suppress noise while preserving edges.
2. Compute horizontal and vertical gradients using the Scharr operator.
3. Form unnormalised normal vectors (n_x, n_y, n_z) by combining gradients with a constant z component.
4. Normalise each vector to unit length, then smooth each component with a Gaussian filter.
5. Add an ambient offset to n_z and apply gamma correction.
6. Linearly scale to 8-bit range, then apply CLAHE for local contrast enhancement.
7. Multiply by 0.6 to moderate influence, and merge with the raw depth and intensity channels into a three-channel DIN image.

This pipeline runs in approximately 0.02 s per frame on a CPU, adding negligible overhead to the real-time system.

4.B.2 Noise Type Details

The noise type implementation specifics are as follows. Basic `Noise` corruptions add zero-mean Gaussian noise with single scale factors per modality: RGB receives channel-wise additive noise in the [0,255] domain with scale proportional to a global intensity parameter; thermal (T24) and UV (U8) use modality-aware scales; DIN employs depth-specific scaling; all outputs are clipped to valid ranges. Advanced `Gaussian` corruptions adapt noise levels to each channel's standard deviation with modality-specific minima to prevent vanishing perturbations (approximately 15 for DIN, 20 for T24, 10 for U8, and 20 for RGB in pixel units). Advanced `SaltPepper` corruptions use modality-specific corruption probabilities (DIN: 0.15×intensity, T24: 0.20×, U8: 0.10×, RGB: 0.10×). Advanced `Speckle` corruptions apply multiplicative noise with modality-specific gains (DIN: 0.4×, T24: 0.6×, U8: 0.3×, RGB: 0.35×), with results clamped to valid ranges.

4.C Ablation Details

Table 4.9: Comprehensive Ablation Analysis Summary

RGB Config	Scenario	Ablation Type	Affected Modality	mIoU	Degradation %	T Gate Mean	UV Gate Mean	T Stage0 Mean	UV Stage0 Mean	T Stage1 Mean	UV Stage1 Mean	T Stage2 Mean	UV Stage2 Mean	T Stage3 Mean	UV Stage3 Mean
RGB1	Full	Baseline	None	84.9	0.0	0.594	0.405	0.999	0.623	0.377	0.993	1.000	0.002	0.000	0.000
RGB1	Drop_DIN	Complete Removal	DIN	27.9	67.1	0.580	0.407	0.998	0.637	0.323	0.990	1.000	0.000	0.000	0.000
RGB1	Drop_RGB	Complete Removal	RGB	20.8	75.5	0.501	0.329	0.999	0.316	0.006	1.000	1.000	0.000	0.000	0.000
RGB1	Drop_T24	Complete Removal	T24	29.1	65.7	0.630	0.405	0.981	0.623	0.537	0.997	1.000	0.001	0.000	0.000
RGB1	Drop_U8	Complete Removal	U8	57.8	31.9	0.600	0.341	0.999	0.362	0.402	1.000	1.000	0.000	0.000	0.000
RGB1	Gaussian_DIN	Gaussian Noise	DIN	82.9	2.4	0.589	0.408	0.999	0.634	0.357	0.996	1.000	0.002	0.000	0.000
RGB1	Gaussian_RGB	Gaussian Noise	RGB	34.2	59.7	0.726	0.416	0.999	0.900	0.906	0.762	1.000	0.001	0.000	0.000
RGB1	Gaussian_T24	Gaussian Noise	T24	49.2	42.1	0.609	0.404	0.993	0.623	0.444	0.992	1.000	0.002	0.000	0.000
RGB1	Gaussian_U8	Gaussian Noise	U8	79.4	6.5	0.584	0.394	0.999	0.576	0.338	0.997	1.000	0.002	0.000	0.000
RGB1	Noise_DIN	Basic Noise	DIN	84.0	1.1	0.589	0.407	0.999	0.633	0.359	0.993	1.000	0.002	0.000	0.000
RGB1	Noise_RGB	Basic Noise	RGB	30.9	63.6	0.733	0.422	0.998	0.922	0.935	0.765	1.000	0.001	0.000	0.000
RGB1	Noise_T24	Basic Noise	T24	43.8	48.5	0.608	0.404	0.985	0.623	0.446	0.993	1.000	0.002	0.000	0.000
RGB1	Noise_U8	Basic Noise	U8	83.3	1.9	0.588	0.392	0.999	0.569	0.351	0.997	1.000	0.003	0.000	0.000
RGB1	SaltPepper_DIN	Salt&Pepper Noise	DIN	82.4	2.9	0.590	0.408	0.999	0.636	0.360	0.995	1.000	0.002	0.000	0.000
RGB1	SaltPepper_RGB	Salt&Pepper Noise	RGB	31.0	63.5	0.712	0.399	0.994	0.862	0.855	0.731	1.000	0.001	0.000	0.000
RGB1	SaltPepper_T24	Salt&Pepper Noise	T24	29.9	64.8	0.610	0.405	0.966	0.623	0.472	0.994	1.000	0.001	0.000	0.000
RGB1	SaltPepper_U8	Salt&Pepper Noise	U8	81.0	4.6	0.585	0.397	0.999	0.586	0.343	0.997	1.000	0.003	0.000	0.000
RGB1	Speckle_DIN	Speckle Noise	DIN	82.4	3.0	0.585	0.408	0.999	0.636	0.341	0.992	1.000	0.002	0.000	0.000
RGB1	Speckle_RGB	Speckle Noise	RGB	81.2	4.4	0.594	0.402	0.999	0.614	0.378	0.992	1.000	0.002	0.000	0.000
RGB1	Speckle_T24	Speckle Noise	T24	29.8	65.0	0.607	0.405	0.962	0.623	0.466	0.995	1.000	0.001	0.000	0.000
RGB1	Speckle_U8	Speckle Noise	U8	84.4	0.6	0.591	0.384	0.999	0.535	0.366	0.998	1.000	0.003	0.000	0.000
RGB3	Full	Baseline	None	88.3	0.0	0.826	0.252	0.999	0.434	0.306	0.574	1.000	0.000	1.000	0.000
RGB3	Drop_DIN	Complete Removal	DIN	58.9	33.3	0.810	0.255	0.999	0.420	0.240	0.600	1.000	0.000	1.000	0.000
RGB3	Drop_RGB	Complete Removal	RGB	8.6	90.2	0.750	0.194	0.999	0.296	0.003	0.478	1.000	0.000	1.000	0.000
RGB3	Drop_T24	Complete Removal	T24	37.1	58.0	0.831	0.256	0.985	0.434	0.340	0.589	1.000	0.000	1.000	0.000
RGB3	Drop_U8	Complete Removal	U8	65.4	25.9	0.827	0.290	0.999	0.189	0.309	0.972	1.000	0.000	1.000	0.000
RGB3	Gaussian_DIN	Gaussian Noise	DIN	87.7	0.7	0.828	0.255	0.999	0.438	0.315	0.580	1.000	0.000	1.000	0.000
RGB3	Gaussian_RGB	Gaussian Noise	RGB	51.0	42.2	0.800	0.285	0.998	0.505	0.202	0.636	1.000	0.000	1.000	0.000
RGB3	Gaussian_T24	Gaussian Noise	T24	63.7	27.8	0.836	0.250	0.991	0.434	0.354	0.564	1.000	0.000	1.000	0.000
RGB3	Gaussian_U8	Gaussian Noise	U8	82.9	6.1	0.820	0.176	0.999	0.363	0.280	0.341	1.000	0.000	1.000	0.000
RGB3	Noise_DIN	Basic Noise	DIN	88.1	0.2	0.826	0.252	0.999	0.433	0.304	0.573	1.000	0.000	1.000	0.000
RGB3	Noise_RGB	Basic Noise	RGB	81.6	7.5	0.816	0.256	0.999	0.433	0.266	0.592	1.000	0.000	1.000	0.000
RGB3	Noise_T24	Basic Noise	T24	56.9	35.5	0.835	0.249	0.986	0.434	0.355	0.563	1.000	0.000	1.000	0.000
RGB3	Noise_U8	Basic Noise	U8	87.1	1.3	0.824	0.223	0.999	0.388	0.297	0.502	1.000	0.000	1.000	0.000
RGB3	SaltPepper_DIN	Salt&Pepper Noise	DIN	87.6	0.7	0.830	0.253	0.999	0.436	0.322	0.575	1.000	0.000	1.000	0.000
RGB3	SaltPepper_RGB	Salt&Pepper Noise	RGB	64.7	26.7	0.812	0.277	0.999	0.488	0.251	0.620	1.000	0.000	1.000	0.000
RGB3	SaltPepper_T24	Salt&Pepper Noise	T24	47.5	46.2	0.833	0.249	0.977	0.434	0.356	0.561	1.000	0.000	1.000	0.000
RGB3	SaltPepper_U8	Salt&Pepper Noise	U8	84.9	3.8	0.822	0.202	0.999	0.388	0.288	0.421	1.000	0.000	1.000	0.000
RGB3	Speckle_DIN	Speckle Noise	DIN	88.1	0.2	0.824	0.251	0.999	0.435	0.298	0.569	1.000	0.000	1.000	0.000
RGB3	Speckle_RGB	Speckle Noise	RGB	71.8	18.7	0.807	0.245	0.999	0.407	0.230	0.571	1.000	0.000	1.000	0.000
RGB3	Speckle_T24	Speckle Noise	T24	45.6	48.4	0.835	0.251	0.972	0.434	0.366	0.569	1.000	0.000	1.000	0.000
RGB3	Speckle_U8	Speckle Noise	U8	88.1	0.2	0.826	0.232	0.999	0.380	0.304	0.550	1.000	0.000	1.000	0.000
RGB5	Full	Baseline	None	84.2	0.0	0.994	0.591	0.998	0.591	0.977	0.735	1.000	0.037	1.000	1.000
RGB5	Drop_DIN	Complete Removal	DIN	25.9	69.2	0.992	0.582	0.998	0.592	0.972	0.736	1.000	0.002	1.000	1.000
RGB5	Drop_RGB	Complete Removal	RGB	13.0	84.6	0.886	0.594	0.994	0.358	0.550	0.951	1.000	0.067	1.000	1.000
RGB5	Drop_T24	Complete Removal	T24	28.4	66.3	0.940	0.595	0.814	0.591	0.946	0.766	1.000	0.023	1.000	1.000
RGB5	Drop_U8	Complete Removal	U8	56.9	32.4	0.993	0.634	0.998	0.459	0.976	1.000	1.000	0.078	1.000	1.000
RGB5	Gaussian_DIN	Gaussian Noise	DIN	82.2	2.4	0.994	0.600	0.998	0.589	0.979	0.754	1.000	0.058	1.000	1.000
RGB5	Gaussian_RGB	Gaussian Noise	RGB	42.8	49.1	0.958	0.588	0.998	0.508	0.835	0.779	1.000	0.065	1.000	1.000
RGB5	Gaussian_T24	Gaussian Noise	T24	55.3	34.3	0.988	0.594	0.978	0.591	0.976	0.745	1.000	0.040	1.000	1.000
RGB5	Gaussian_U8	Gaussian Noise	U8	79.5	5.6	0.994	0.583	0.998	0.782	0.978	0.536	1.000	0.013	1.000	1.000
RGB5	Noise_DIN	Basic Noise	DIN	83.4	1.0	0.994	0.590	0.998	0.584	0.977	0.737	1.000	0.039	1.000	1.000
RGB5	Noise_RGB	Basic Noise	RGB	46.7	44.6	0.966	0.631	0.998	0.563	0.867	0.810	1.000	0.152	1.000	1.000
RGB5	Noise_T24	Basic Noise	T24	49.4	41.3	0.983	0.595	0.962	0.591	0.968	0.748	1.000	0.040	1.000	1.000
RGB5	Noise_U8	Basic Noise	U8	83.3	1.0	0.994	0.600	0.998	0.710	0.978	0.666	1.000	0.022	1.000	1.000
RGB5	SaltPepper_DIN	Salt&Pepper Noise	DIN	81.7	3.0	0.994	0.598	0.998	0.581	0.979	0.746	1.000	0.063	1.000	1.000
RGB5	SaltPepper_RGB	Salt&Pepper Noise	RGB	39.1	53.6	0.913	0.632	0.998	0.583	0.656	0.805	1.000	0.142	1.000	1.000
RGB5	SaltPepper_T24	Salt&Pepper Noise	T24	38.0	54.9	0.970	0.597	0.931	0.591	0.947	0.755	1.000	0.040	1.000	1.000
RGB5	SaltPepper_U8	Salt&Pepper Noise	U8	80.7	4.1	0.994	0.595	0.998	0.755	0.978	0.609	1.000	0.016	1.000	1.000
RGB5	Speckle_DIN	Speckle Noise	DIN	82.1	2.5	0.994	0.594	0.998	0.585	0.977	0.741	1.000	0.049	1.000	1.000
RGB5	Speckle_RGB	Speckle Noise	RGB	16.9	80.0	0.880	0.665	0.997	0.577	0.524	0.867	1.000	0.216	1.000	1.000
RGB5	Speckle_T24	Speckle Noise	T24	33.8	59.9	0.971	0.596	0.935	0.591	0.950	0.754	1.000	0.038	1.000	1.000
RGB5	Speckle_U8	Speckle Noise	U8	83.9	0.3	0.994	0.596	0.998	0.661	0.977	0.696	1.000	0.027	1.000	1.000

Bibliography

[1] Martin Brenner, Napoleon H. Reyes, Teo Susnjak, and Andre L.C. Barczak. Mm5: Multimodal image capture and dataset generation for rgb, depth, thermal, uv, and nir. *Information Fusion*, 126:103516, 2026.

[2] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023.

[3] Jiyuan Qiu, Chen Jiang, and Haowen Wang. ETFormer: An Efficient Transformer Based on Multimodal Hybrid Fusion and Representation Learning for RGB-D-T Salient Object Detection. *IEEE Signal Processing Letters*, 31:2928–2932, 2024.

[4] Christian Stippel, Thomas Heitzinger, and Martin Kampel. A trimodal dataset: Rgb, thermal, and depth for human segmentation and temporal action detection. In *DAGM German Conference on Pattern Recognition*, pages 18–33. Springer, 2023.

[5] Nianchang Huang, Yang Yang, Ruida Xi, Qiang Zhang, Jungong Han, and Jin Huang. Salient Object Detection From Arbitrary Modalities. *arXiv preprint arXiv:2405.03352*, 2024. Under review.

[6] Kechen Song, Han Wang, Ying Zhao, Liming Huang, Hongwen Dong, and Yunhui Yan. Lightweight multi-level feature difference fusion network for RGB-D-T salient object detection. *Journal of King Saud University - Computer and Information Sciences*, 35(10):101702, 2023.

[7] Liuxin Bao, Xiaofei Zhou, Xiankai Lu, Yaoqi Sun, Haibing Yin, Zhenghui Hu, Jiyong Zhang, and Chenggang Yan. Quality-aware selective fusion network for vdt salient object detection. *IEEE Transactions on Image Processing*, 33:3212–3226, 2024.

[8] M. Brenner, N. H. Reyes, T. Susnjak, and A. L. C. Barczak. RGB-D and thermal sensor fusion: A systematic literature review. *IEEE Access*, 11:102667–102685, 2023.

[9] Kechen Song, Jie Wang, Yanqi Bao, Liming Huang, and Yunhui Yan. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. *IEEE/ASME Transactions on Mechatronics*, 28(3):1558–1569, 2022.

[10] Hongwei Wen, Kechen Song, Liming Huang, Han Wang, Junyi Wang, and Yunhui Yan. Hierarchical two-stage modal fusion for triple-modality salient object detection. *Measurement*, 218:113180, 2023.

[11] Bin Wan, Xiaofei Zhou, Yaoqi Sun, Zunjie Zhu, Hongkui Wang, Chenggang Yan, et al. Tmnet: Triple-modal interaction encoder and multi-scale fusion decoder network for vdt salient object detection. *Pattern Recognition*, 147:110074, 2024.

[12] Ahmet Ozcan and Omer Cetin. A Novel Fusion Method With Thermal and RGB-D Sensor Data for Human Detection. *IEEE Access*, 10:66831–66840, 2022.

[13] Martin Brenner, Napoleon Reyes, Teo Susnjak, and Andre Barczak. MM5: Multimodal Image Dataset, 2025. Dataset.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[15] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 933–941, 2017.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[17] Javier Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[18] Yuanzhouhan Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3029–3037, 2017.

[19] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020.

[20] Etienne Balit and Amine Chadli. Gmfnet: Gated multimodal fusion network for visible-thermal semantic segmentation. In *Proceedings 16th the European Conference on Computer Vision*, pages 1–4, 2020.

[21] Yongjie Guo, Feng Wang, Yuming Xiang, and Hongjian You. Dgfnet: dual gate fusion network for land cover classification in very high-resolution images. *Remote Sensing*, 13(18):3755, 2021.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[23] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 2002.

[24] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[25] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*. OpenReview.net, 2019. Presented at the 7th International Conference on Learning Representations (ICLR 2019).

[27] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. arXiv:1905.11946.

[28] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, pages 7262–7272, 2021. arXiv:2105.05633.

[29] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[30] Mark G. Vangel. Confidence intervals for a normal coefficient of variation. *The American Statistician*, 50(1):21–26, 1996.

Chapter 5

Decoder-Level Fusion with Dedicated Thermal/UV Heads

The encoder-level fusion architecture presented in Chapter 4 achieved strong segmentation accuracy at real-time throughput under aligned conditions, but revealed a critical limitation: learnt gates specialise in training distributions and fail to adapt when modalities are degraded or removed, with complete RGB loss triggering up to 90% performance degradation. This chapter investigates whether decoder-level fusion can provide greater robustness to such failures. Using the unaligned subset of MM5 introduced in Chapter 3—comprising raw lens-distorted thermal and UV imagery with MAR-reprojected annotations—this chapter develops CMAG (Cross-Modal Attention with Gated Residuals) and five architecturally matched decoder-level baselines. The systematic comparison between encoder-level (Chapter 4) and decoder-level fusion quantifies a fundamental trade-off: encoder fusion achieves higher peak accuracy and throughput, while decoder fusion demonstrates improved resilience to modality dropout and sensor noise, and exhibits robust tolerance to spatial misalignment—offering deployment guidance for scenarios where sensor reliability or geometric calibration cannot be guaranteed.

This chapter is based on:

Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2025). Pre-Logit Decoder Fusion for Five-Modality Segmentation with Unaligned T/UV Auxiliaries.

Status: The manuscript was submitted to Information Fusion.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Martin Brenner
Name and title of main supervisor:	Dr Napoleon Reyes
In which chapter is the manuscript/published work?	5

Describe the contribution that the student and members of the supervisory team have made to the manuscript/published work:[1]

The candidate was the main contributor of this work, and has done the literature review, experiments, and drafted the manuscript. The final draft was completed with the suggestions from the co-authors.

Please select one of the following three options:

○ **The manuscript/published work is published or in press**

Please provide the full reference of the research output:

◉ **The manuscript is currently under review for publication**

Please provide the name of the journal:

Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2025). Pre-Logit Decoder Fusion for Five-Modality Segmentation with Unaligned T/UV Auxiliaries. Information Fusion.

○ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Digitally signed by Martin Brenner DN: cn=Martin Brenner, c=NZ, email=mb@fisag.co.nz Reason: I agree to specified portions of this document Location: Auckland Date: 2025.11.25 20:47:40 +13'00'	Main supervisor's signature:	Napoleon Reyes Digitally signed by Napoleon Reyes Date: 2025.12.01 17:09:54 +13'00'

This form should be placed at the beginning of each relevant thesis chapter.

[1] Refer to the Massey University Publishing and Authorship guidelines (OneMassey for staff, Stream for students) and/ or Contributor Roles Taxonomy (CRediT) guidelines for guidance.

5.1 Abstract

We investigate decoder-level multimodal fusion for semantic segmentation with unaligned modalities (RGB+DIN(depth-intensity-normals), thermal, and ultraviolet (UV)). We introduce Cross-Modal Attention with Gated Residuals (CMAG), a hybrid module operating at the pre-logit stage that combines two complementary pathways: Global Context Modality Attention (GCMA), which establishes soft correspondences between thermal/UV and RGB+DIN features, and sigmoid-gated (SIG) residuals that inject per-pixel corrections from auxiliary modalities. Independent decoders generate quarter-resolution pre-logits per modality, preserving modularity while enabling robust handling of missing inputs, thereby eliminating explicit calibration requirements. We implement five decoder-level baselines, adapting established fusion paradigms to contextualise CMAG's performance.
On the MM5 dataset, CMAG achieves 84.18% mIoU across lighting conditions (underexposed/ideal/overexposed: 82.54%/87.61%/82.38%), outperforming attention-only GCMA (80.49%/86.72%/78.03%). Ablations reveal the importance of hierarchical modality: RGB and DIN removal cause severe degradation (59.50 pp and 49.61 pp, respectively), while thermal and UV provide specialised cues (24.62pp and 16.82pp losses). Spatial misalignment proves substantially less damaging than modality removal (20-pixel shifts: 2.61 pp vs 37.64 pp for drops), validating decoder fusion's alignment tolerance. Architectural comparison reveals distinct robustness profiles: CMAG maximises clean-data accuracy but shows elevated noise sensitivity (12.93 pp mean degradation), adapted Multimodal Transfer Module (PL-MMTM) achieves superior modality-drop robustness (31.82 pp), and adapted Recurrent Attention U-Net (PL-R2AU) demonstrates best noise resilience (8.80 pp). Comparison with encoder-level fusion (GF-Net) reveals fundamental trade-offs: encoder fusion achieves +1.34 pp (vs CMAG) accuracy and about 2× throughput, but suffers severe sensor-failure degradation, while decoder fusion prioritises robustness through late integration. These findings establish decoder-level fusion as viable for unaligned multimodal segmentation when robustness outweighs peak performance.

5.2 Introduction

Robust semantic segmentation in field robotics, industrial inspection, and agriculture requires tolerance to fluctuating illumination, reflective or transparent materials, occlusions, and camouflage-like textures. Single-stream RGB systems are brittle under such conditions. In contrast, complementary sensors provide cues that compensate for specific weaknesses: depth supplies geometry independent of colour, thermal imaging highlights temperature-emissive regions, near-infrared intensity (NIR) broadens the dynamic range and extends the observable spectrum, and ultraviolet (UV) reveals fluorescence and sub-visible surface structure [1]. Harnessing this heterogeneity promises finer delineation and more reliable decisions across domestic, industrial, and agricultural settings.
In practical capture rigs, not all streams are geometrically compatible. In our setup, RGB-D-NIR are inherently co-registered by the factory-calibrated RGB-D sensor, whereas UV and thermal (LWIR) use different optics, exhibit lens distortion, and are unaligned with respect to RGB-D-NIR. Early (data-level) and intermediate (feature-level) fusion strategies typically assume either prior geometric registration or rely on learnt, in-network rectification. Examples include CMX's Cross-Modal Feature Rectification Module (FRM) [2] and trimodal encoder attention in ETFormer [3], which adapt one stream using another before mixing. These mechanisms increase complexity and can remain brittle under residual misalignment or viewpoint change, which partly explains the prevalence of two-stream pairings (RGB-D, RGB-T) and the limited scalability to larger, heterogeneous sensor sets.
We propose CMAG (Cross-Modal Attention with Gated Residuals), a decoder-level fusion module that integrates unaligned modalities at the pre-logit stage through two complementary mechanisms. The primary path (RGB augmented with DIN (Depth, Intensity, and Normals [1]) forms the base representation, while LWIR thermal (T24) and UV (U8) are processed through separate encoder-decoder branches, preserving their native geometry. At the decoder's pre-logit stage, GCMA (Global Context Modality Attention) performs efficient cross-modal attention by querying thermal/UV features with RGB-DIN representations using downsampled tokens, extracting global context without requiring explicit spatial calibration. CMAG then applies sigmoid-gated (SIG) residuals, injecting per-pixel auxiliary modality corrections directly in feature space before final classification through a lightweight 1×1 convolution. Unlike encoder-level fusion, which presupposes spatial alignment, CMAG operates directly on unaligned sensor streams and remains compatible with optional alignment modules when available (Figure 5.1). Additionally, we introduce MWPA (Modality-Wise Parallel Attention). This computationally more efficient alternative employs parallel channel and

spatial attention mechanisms for modality-selective fusion, providing a lightweight baseline for comparative analysis of decoder-level fusion strategies.

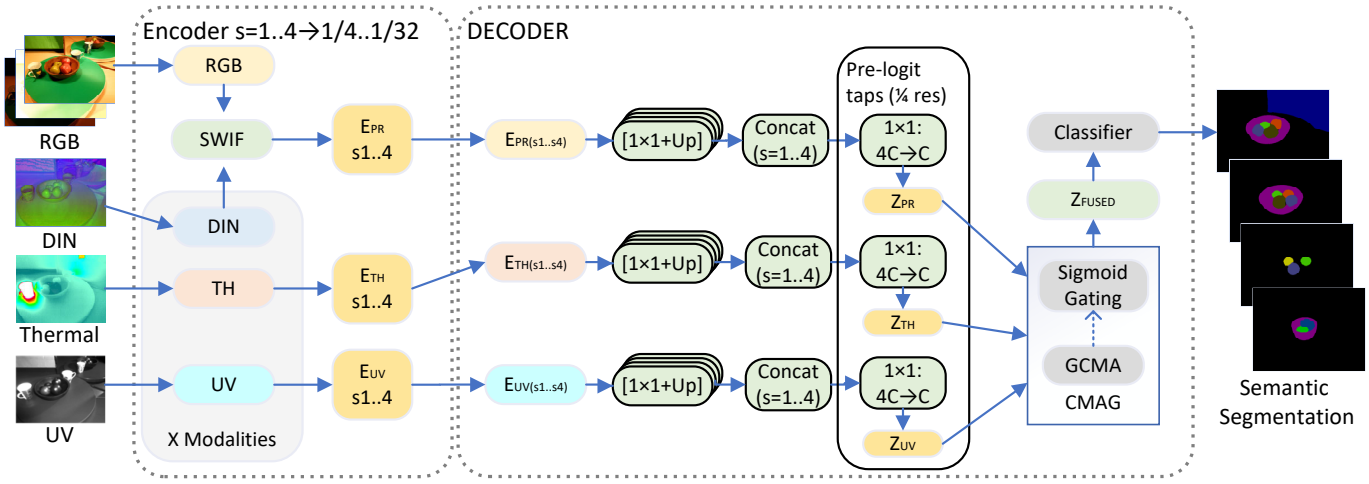


Figure 5.1: Overview of CMAG with three streams (RGB+DIN, LWIR, UV). At each encoder stage $s \in \{1, 2, 3, 4\}$, the primary path applies SWIF [4] to enhance RGB with DIN, yielding $E_{\mathrm{PR}}^{(s)}$; thermal and UV produce $E_{\mathrm{TH}}^{(s)}$ and $E_{\mathrm{UV}}^{(s)}$ while preserving native geometry. Each decoder consumes its stage bundle $\{E^{(s)}\}$, applies $[\ 1{\times}1 + \mathrm{Up}{\to}\frac{1}{4}\ ]$ per stage, Concats over $s{=}1..4$, then reduces channels with $1{\times}1 : 4C{\to}C$ (SE optional) to form the pre-logits $Z_{\mathrm{PR}}, Z_{\mathrm{TH}}, Z_{\mathrm{UV}}$ at $\frac{1}{4}$ resolution. CMAG fuses these pre-logits: GCMA attends from Z_{PR} to $(Z_{\mathrm{TH}}, Z_{\mathrm{UV}})$ to produce F_{gcma}, and sigmoid-gated residuals add thermal/UV contributions to obtain $Z_{\mathrm{fused}} = F_{\mathrm{gcma}} + r_{\mathrm{TH}} + r_{\mathrm{UV}}$. A $1{\times}1$ classifier and $\times 4$ upsampling yield fused logits at $H{\times}W$ without explicit geometric warping of LWIR/UV. Here, s indexes resolutions from $\frac{1}{4}$ to $\frac{1}{32}$; C denotes the decoder channel width.

To establish comprehensive decoder-level benchmarks and contextualise CMAG's performance, we adapt established fusion paradigms, PL-MMTM and PL-R2AU, to our alignment-free framework. PL-MMTM extends the channel-wise squeeze-and-excitation mechanism [5] to pre-logit features, enabling cross-modal salience transfer without spatial correspondence. PL-R2AU adapts recurrent attention gates [6] for consistent spatial focus across modalities. We evaluate GCMA and sigmoid gating as standalone modules to isolate the constituent mechanisms of CMAG. This unified framework, comprising six decoder-level variants, enables a systematic assessment of fusion complexity versus performance trade-offs, ranging from lightweight gating (PL-SIG) and parallel attention (MWPA) to channel modulation (PL-MMTM), recurrent spatial attention (PL-R2AU), and global cross-modal attention (GCMA, CMAG). To support training and evaluation in this alignment-free regime, we curate a new MM5 subset comprising raw, unaligned RGB-D-NIR-T-UV imagery. Thermal and UV undergo only coarse preprocessing to establish overlapping fields of view, followed by random crops that mimic realistic misalignment and background variation (Section 5.4). This subset complements the aligned MM5 release [1].

5.2.1 Key Contributions

The main contributions of this work are:

1. We introduce CMAG, a decoder-level fusion module that integrates unaligned thermal and UV streams into an RGB-Depth-Intensity-Normals backbone via global cross-modal attention and sigmoid-gated residuals, achieving alignment-tolerant fusion without explicit geometric calibration.
2. We design a family of decoder-level fusion baselines by adapting established encoder/feature-level mechanisms (MMTM, R2AU, GF-Net style sigmoid gating and Modality-Wise Parallel Attention) to the pre-logit stage, enabling controlled, like-for-like comparisons of gating versus attention under a shared backbone and training protocol.
3. We conduct an extensive robustness characterisation, systematically evaluating all six decoder-level fusion architectures across modality dropout, spatial misalignment, and sensor noise injection, with a systematic comparison against encoder-level fusion to quantify the impact of fusion stage choice on accuracy-robustness trade-offs.
4. We curate and release an unaligned MM5 subset with raw, lens-distorted thermal and UV imagery, together with code and trained weights for CMAG and all decoder-level baselines, providing a reproducible benchmark for

multimodal segmentation with misaligned auxiliary sensors.

5.3 Related Work

Decoder-Level Fusion Methods Decoder-level fusion integrates modalities during the upsampling phase, before final classification. Three architectural patterns dominate this space. Hyper-fusion decoders aggregate multi-scale features from separate encoders: OctopusNet merges per-modality pyramids at each decode stage [7]. Multi-branch decoders maintain separate pathways: Mirror U-Net pairs modality-specific decoders with an auxiliary multimodal decoder and consistency objectives [8]. Central fusion decoders route all modalities through shared layers: SGFNet derives semantic guidance maps to reweight features [9], MEFNet employs modality experts [10], and GMFNet applies pixel-wise gates within U-shaped decoders [11].
Whilst these architectures demonstrate the effectiveness of feature-level integration within decoder structures, recent work has explored the extreme of purely decision-level approaches. LF-DLM reports approximately 0.3% mIoU gains by fusing modality-specific logits while keeping encoders and decoders completely independent [12]. This minimal improvement establishes a crucial baseline, highlighting that the marginal benefits of pure late fusion motivate exploration of more sophisticated cross-modal mechanisms that can capture richer inter-modality relationships without sacrificing the modularity advantages of decoder-level architectures.

Cross-Modal Attention for Fusion Building upon the insights from decoder-level fusion limitations, cross-modal attention mechanisms have emerged as a powerful paradigm for integrating heterogeneous streams without requiring strict spatial alignment. CMAF-Net embeds cross-modal attention within a multi-encoder 3D U-Net, learning modality-invariant latents from incomplete multi-sequence MRI [13]. The CMAF block achieves strong Dice scores on BraTS 2020 under conditions of missing modality, demonstrating robustness to incomplete data. CMNeXt scales to arbitrary modalities on DeLiVER through late, lightweight attention-based integration with minimal per-modality parameters [14]. CANet employs bidirectional co-attention between RGB and depth features [15], while UCTNet incorporates uncertainty-aware cross-modal transformers [16]. These attention-based approaches offer dynamic, learnable mechanisms for modality interaction that adapt to input characteristics, addressing the limitations of static fusion rules observed in pure decision-level methods.

Attention-Based Feature Refinement Whilst cross-modal attention facilitates inter-modality exchange, complementary research has focused on refining individual modality representations through channel and spatial recalibration before fusion occurs. CBAM applies sequential attention [17], while BAM employs parallel branches [18]; both operate on unimodal features to enhance their discriminative power. In multimodal contexts, this intra-modal refinement proves particularly valuable: MEFNet introduces modality-specific expert networks with attention-based reweighting [10], and TriFuse applies tri-attention across spatial, channel, and modality dimensions [19]. These refinement strategies recognise that optimal fusion requires not only effective cross-modal interaction but also maximally informative individual modality representations. Unlike methods that process modalities independently or via cross-attention [2], our MWPA bridges these paradigms by deriving per-modality attention weights from concatenated representations, enabling simultaneous recalibration of both channels and spatial dimensions before fusion.

Gating Mechanisms Whilst attention mechanisms recalibrate features, learnt gating directly assigns per-pixel or per-channel reliability weights to modulate modality contributions. MMTM implements channel-wise squeeze-and-excitation for cross-modal salience transfer [5]. SSMA demonstrates sigmoid-style recalibration for adaptive feature scaling [20]. GMFNet embeds pixel-wise gates in U-shaped decoders [11]. DGFNet's dual-gate design merges spatial detail with semantics during decoding [21]. R2AU-Net incorporates recurrent attention gates for consistent spatial focus [6].
Quality-aware gating extends basic reliability weighting. QSF-Net estimates quality maps for trimodal visible-depth-thermal integration [22]. MGFNet applies lightweight gating for optical-SAR fusion under cross-sensor noise [23]. These mechanisms prove valuable when modality quality varies or when inputs misalign.

Feature-Level Fusion Baselines In contrast to decoder-level gating approaches, encoder-level fusion methods integrate modalities earlier in the network, achieving strong performance when inputs align. CMX's cross-modal

feature rectification module (FRM) calibrates features before mixing [2]. ETFormer demonstrates single-encoder multimodal attention for RGB-D-T through task-specific pretraining [3]. GF-Net performs early RGB enhancement using SWIF and per-pixel sigmoid gating at encoder stages [4] for semantic segmentation using five modality fusion. These methods assume or restore spatial correspondence before decoding, limiting their applicability to unaligned sensors.

5.3.1 Alignment Handling in Decoder Fusion

The alignment dependency of encoder-level methods motivates decoder fusion approaches that accommodate misregistration through soft alignment via cross-modal attention or fusion at coarser scales, where parallax reduces [24, 25]. When stronger consistency is required, feature rectification precedes late fusion (e.g., CMX's FRM [2]). Performance differences between fusion points depend on the dataset and task rather than a universal ranking [2, 25]. Recent methods directly target misalignment during decoding. Project-and-Fuse uses texture-prior guidance to mitigate biassed assignment in unaligned streams without pixel-exact registration, optimising a divergence-based loss to discourage degenerate assignments [26]. LMFNet employs lightweight transformer decoders that attend to heterogeneous cues across scales, while keeping compute manageable for high-resolution inputs [27].

5.3.2 Uncertainty and Dynamic Reliability in Decoder Fusion

Beyond handling spatial misalignment, dynamic reliability modelling at the decoder extends basic gating mechanisms. UDFNet couples uncertainty estimation with dynamic fusion, reporting strong accuracies across Berlin, Augsburg, MUUFL and Trento benchmarks, outperforming prior decoder-only fusion by sizeable margins [28]. However, the strongest results in overhead remote sensing often come from hybrid pipelines blending feature-level and decision-level fusion (e.g., MCAM/CPS modules with decision fusion in TCPSNet, prototype compensation in PICNet), indicating room for improvement in decoder-only strategies [29].

5.3.3 Positioning of CMAG

CMAG addresses decoder-level fusion for unaligned, heterogeneous sensors. Existing decoder-level methods assume aligned inputs [7, 11] or use coarse decision-level fusion [12]. Encoder-level approaches [2, 3] typically operate on stereo-calibrated, pre-registered datasets and employ feature rectification to refine alignment at the encoder level. Explicit spatial warping via parametric transformations [30, 31] can accommodate geometric distortions but introduces computational overhead. In contrast, CMAG operates directly on unaligned auxiliary modalities at the decoder level, leveraging cross-modal attention for alignment-tolerant fusion without explicit spatial transformation.
Architecturally, CMAG performs single-stage fusion at the pre-logit level after multi-scale feature aggregation. Global Context Modality Attention (GCMA) establishes soft correspondence by attending to pooled modality representations rather than dense spatial tokens, accommodating misalignment with reduced complexity. Sigmoid-gated residuals add fine-grained spatial corrections. This design contrasts with per-stage fusion [11], concatenation-based methods [9, 10], and full spatial attention approaches [13].
Separate per-modality decoder heads provide granular supervision and enable more graceful degradation under missing inputs, while learnt gate maps offer spatial interpretability of modality contributions. This modularity distinguishes CMAG from single-head architectures and channel-wise methods [5].
Table 5.1 summarises the supervision strategies and architectural choices across these methods, highlighting the diversity of approaches to modality-specific training and inference.

5.4 MM5 Dataset

The MM5 dataset [1], introduced in our prior work, was developed to address persistent limitations in existing multimodal segmentation benchmarks, most notably their constrained modality diversity, lack of raw sensor fidelity, and absence of unaligned annotation protocols. MM5 integrates five distinct imaging modalities: RGB, depth (D), thermal (T), ultraviolet (UV), and near-infrared (NIR) within a unified acquisition and annotation pipeline. The dataset was constructed using a custom acquisition rig built from off-the-shelf RGB-D sensors and complemented by thermographic and UV imaging systems. Unlike existing datasets that prioritise pre-registered inputs, MM5 provides

Table 5.1: Overview of modality-specific heads and GT usage in representative multimodal fusion methods. "Mod.-specific heads" = separate output heads per modality during training; "Mod.-specific GT" = different targets per modality.

Method	**Year**	**Modalities**	**Fusion category**	**Mod.-specific heads**	**Mod.-specific GT**	**Supervision summary**	**Inference head(s)**
Mirror U-Net [8]	2023	PET+CT; MRI sequences	Decoder feature (multi-branch)	Yes	Yes	Modality-specific decoders with task-tailored supervision; auxiliary multimodal decoder.	Fused/combined
GMFNet [11]	2020	RGB+T	Decoder feature (central+lateral)	Yes	No	Two lateral unimodal + one central multimodal U-Net; shared semantic GT.	Central decoder
SGFNet [9]	2023	RGB+T	Decoder feature (central)	No	No	Semantic guidance maps reweight decoder features	single
MEFNet [10]	2023	RGB+Thermal	Decoder feature (experts)	No	No	Modality experts aggregated in shared decoder	single
DGFNet [21]	2021	RGB+T	Decoder feature (dual-gate)	No	No	Positional and filter gates merge spatial/channel attention	single
R2AU-Net [6]	2021	RGB+T	Decoder feature (recurrent attention)	No	No	Recurrent attention gates for consistent spatial focus	single
OctopusNet [7]	2019	Multi-contrast MRI	Decoder feature (hyper-fusion)	No	No	Separate encoders; hyper-fusion decoder with multi-scale integration.	single
CMAF-Net [13]	2024	MRI (T1, T1ce, T2, FLAIR)	Decoder feature (attention)	No	No	Cross-modal attention fusion in 3D U-Net; handles missing modalities.	single
CMNeXt [14]	2023	Arbitrary modalities	Decoder attention	No	No	Self-query hub selects informative tokens per modality	single
QSF-Net [22]	2024	RGB+D+T	Hybrid (multi-stage)	Yes (stage-1)	No	Stage-wise: saliency, quality maps, fused saliency + edge.	single
UDFNet [28]	2025	HSI/SAR/LiDAR	Decoder feature (uncertainty)	No	No	Uncertainty-aware dynamic fusion at decode time	single
Project-and-Fuse [26]	2025	RGB+D (unaligned)	Alignment-free decoder	No	No	Texture-prior guided fusion; handles unaligned inputs	single
LF-DLM [12]	2024	VHR Aerial+Sentinel-2	Decision-level	Yes (per branch)	No	Per-branch probabilities fused by weighted geometric mean.	Combined outputs
CMAG (ours)	2025	RGB-Depth-NIR + Thermal + UV	Decoder feature (central fusion; single-pass pre-logit attention + gated residuals)	Yes	Yes	Separate heads for primary, thermal and UV plus a fused head; operates on unaligned, lens-distorted UV/LWIR without explicit calibration.	single (auxiliary feature paths active)

Note: For CMAG, thermal and UV decoder branches are executed at inference to produce pre-logit features consumed by the fusion module; their per-modality logits are not used for the final prediction (unless reported for diagnostics).

both geometrically aligned and unaligned variants of each scene, along with pixel-level annotations. This dual-format design supports research on both feature-level fusion and late and decoder-level fusion strategies, without requiring spatial registration.

Each scene in the MM5 dataset contains a varied selection of objects, encompassing fresh produce, plastic replicas, and partially decayed items. The scenes are captured under diverse illumination conditions, including shadows, underexposure, and saturation, ensuring that each modality offers distinct and complementary semantic information. In this work, MM5 serves as a critical resource, facilitating our exploration of decoder-level fusion strategies, particularly in scenarios characterised by imperfect or absent spatial alignment. Moreover, MM5 supports modality-specific supervision through the provision of independent ground truth annotations.

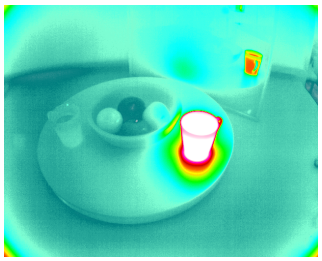
For our experiments on decoder-level fusion, we utilise the unaligned UV and thermal images, which inherently exhibit lens distortion, as illustrated in Figure 5.2. Conversely, depth and NIR images are spatially registered with RGB through the RGB-D sensor.

Following the methodology established in our previous work, GF-Net [4], our analysis focuses specifically on underexposed, well-exposed, and overexposed RGB conditions to enable a direct and consistent comparison.

Due to variations in the image formats across modalities, we identified an approximate region of interest and uniformly cropped all images to a common maximum overlay area measuring 800 pixels in width and 600 pixels in height, as depicted in Figure 5.3. Subsequently, all labelled images were processed and assigned sequential numbering starting

(a) Underexposed RGB (RGB1)

(b) Overexposed RGB (RGB5)

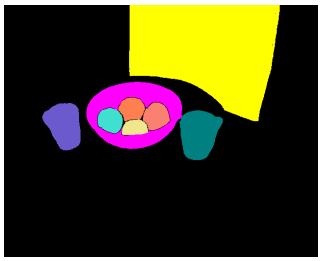


(g) Thermal (T24)

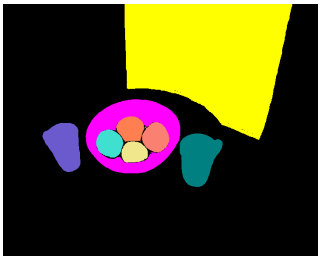
(h) Ultraviolet (U8)

(c) Infrared intensity

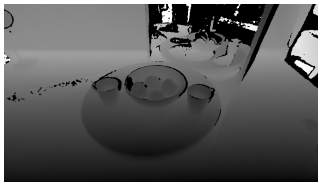
(d) Ideal RGB (RGB3)



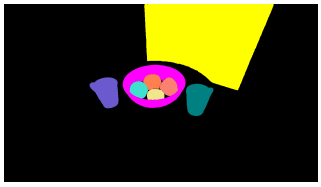
(i) Thermal ground truth mask



(j) UV ground truth mask



(e) Depth



(f) RGB ground truth mask

Figure 5.2: MM5 sample unaligned image subset for frame 544

from 1, analogous to the aligned MM5 dataset, albeit with slight adjustments to folder naming conventions to avoid ambiguity. The class IDs and image IDs remain consistent with the aligned dataset.
Based on the *MM5_RAW_CROPPED* dataset, we created an additional dataset variant with reduced dimensions of 640 pixels in width and 480 pixels in height. To enhance background diversity and facilitate experiments involving camera misalignment, the cropping window position was varied randomly within the original frames. This approach enabled the generation of modality-specific pixel shifts, simulating realistic scenarios of camera displacement. Both datasets will be publicly released alongside the original MM5 dataset as subfolders *MM5_RAW_CROPPED* and *MM5_RAW_CRP640*, the latter includes a metadata file detailing the crop coordinates for each image.

5.4.1 Training and Evaluation Data

For comparability with previous work, we adopted the standard class-wise train-evaluation split provided by the MM5 dataset [1], as detailed in GF-Net [4]. This split maintains an approximate stratification, allocating around 75-80% of the class images to training, with the remaining 20-25% reserved for evaluation. Since the dataset contains mainly mixed scenes, the distribution per class varies. However, the dataset exhibits significant class imbalance, with dominant categories such as Lemon and Mandarin having over a hundred annotated object instances, while others, such as Mandarin Peel and Kettle, are limited to a dozen or fewer examples. Certain classes, like Mandarin Peel, are particularly underrepresented, with evaluation sets as small as three images, while composite scenes introduce additional imbalances. To enhance training effectiveness and mitigate this imbalance, we identified scenes containing underrepresented classes and increased their frequency during training. This was achieved by applying diverse data augmentations, such as zoom, rotation, and flipping. Additionally, the dataset's long-tail class distribution provides an opportunity to evaluate and improve model robustness and generalisation across both rare and frequent classes alike. Further details on the exact composition, challenges, and rationale behind the dataset splits are provided in the original MM5 [1] and GF-Net [4] papers.

5.5 Proposed Methods

We adopt a decoder-centric fusion framework that performs efficient single-stage fusion at the pre-logit level. CMAG integrates two complementary mechanisms: Global Context Modality Attention (GCMA) for cross-modal feature exchange and per-pixel sigmoid gating for fine-grained spatial refinement. This design operates directly on

(a) Thermal image on RGB canvas (b) UV8 image on RGB canvas (c) RGB image

(d) Overlapped RGB, UV (blue frame), and thermal (red frame) (e) Overlapped and cropped RGB, UV, and thermal

Figure 5.3: Overlap and cropping process for frame 544 of the MM5 raw data.

unaligned sensor streams, establishing soft correspondence through learnt attention rather than explicit geometric alignment.

5.5.1 Architecture Overview

The network comprises three coordinated components:

1. **Primary RGB+DIN Stream (PR):** A MiT-B0 encoder enhanced at each stage by Stage-Wise Intensity Fusion (SWIF) [4], which injects pre-computed DIN composites (depth, infrared intensity, surface normals) [1] via lightweight residual blocks for lighting resilience.
2. **Auxiliary Thermal Stream (TH):** An independent MiT-B0 encoder processes long-wave infrared (LWIR) thermal imagery, preserving native geometry without spatial rectification.
3. **Auxiliary Ultraviolet Stream (UV):** An independent MiT-B0 encoder processes ultraviolet imagery, maintaining native geometry without spatial alignment.
4. **Pre-logit CMAG Fusion:** SegFormer-style MLP decoders applied to each stream produce $1/4$-resolution pre-logit features ($Z_{\mathrm{PR}}, Z_{\mathrm{TH}}, Z_{\mathrm{UV}}$). CMAG fuses these via global context attention and sigmoid-gated residuals to generate the final prediction.

Throughout this paper, we use the notation PR (Primary), TH (Thermal), and UV (Ultraviolet) to denote these three modality streams in equations and diagrams.

5.5.2 Decoder and Pre-logit Assembly

Multi-scale integration occurs within each stream's decoder. Each SegFormer-style MLP head projects its four encoder stages to a common embedding dimension C, upsamples them to a unified $1/4$ scale, and aggregates via element-wise summation after 1×1 projection. This produces a single pre-logit feature tensor per stream:

$$Z_{\mathrm{PR}},\ Z_{\mathrm{TH}},\ Z_{\mathrm{UV}} \in \mathbb{R}^{B\times C\times \frac{H}{4}\times \frac{W}{4}}, \tag{5.1}$$

where $C{=}512$ channels encode hierarchical features. Figure 5.4 illustrates this process.

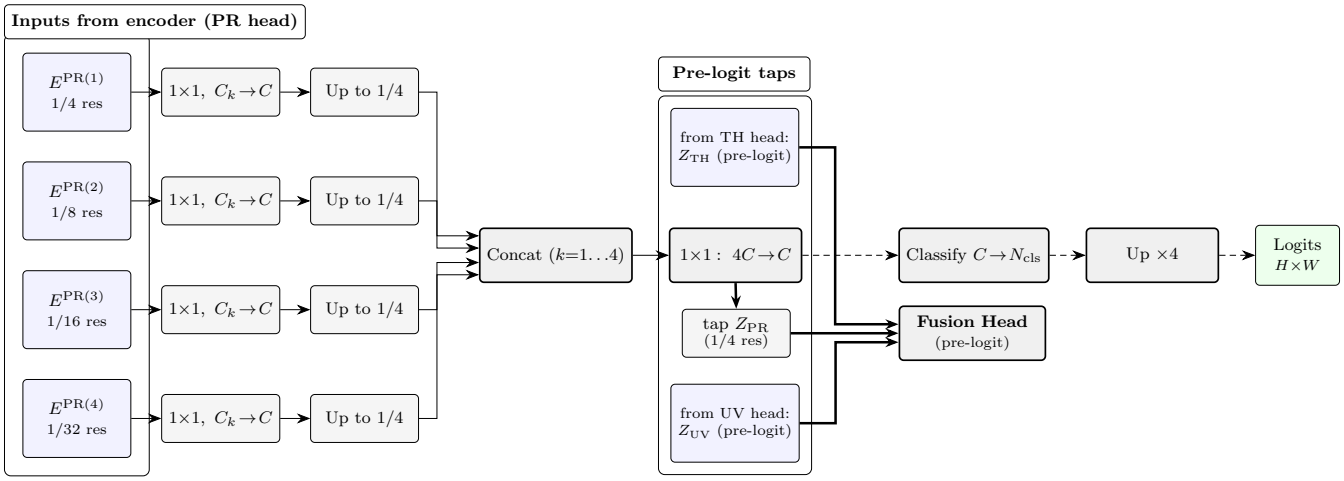


Figure 5.4: SegFormer-style MLP decoder for primary stream. Multi-scale encoder features ($E^{(0)}$ through $E^{(3)}$) are projected to C channels, upsampled to $H/4 \times W/4$, and summed to produce pre-logit features Z_{PR}. *Notation:* res = spatial resolution; C_k = stage width; C = decoder width; N_{cls} = number of classes. The dotted line indicates the single-head classification path.

5.5.3 Pre-logit CMAG Fusion

CMAG fuses the three pre-logit tensors through two sequential stages: global context modality attention (GCMA) followed by sigmoid-gated residuals (SIG).

Global Context Modality Attention (GCMA)

GCMA establishes cross-modal correspondence by operating at the modality level rather than the spatial-token level. Each pre-logit tensor is first lightly enhanced (Conv+Norm) and globally pooled to produce a modality context vector:

$$c_i \;=\; \mathrm{GAP}(\mathrm{Enhance}(Z_i)) \;\in\; \mathbb{R}^{B\times C}, \quad i \in \{\mathrm{PR}, \mathrm{TH}, \mathrm{UV}\}. \tag{5.2}$$

Global Average Pooling (GAP). For $Z \in \mathbb{R}^{B\times C\times H\times W}$,

$$\mathrm{GAP}(Z)_{b,c} \;=\; \frac{1}{HW}\sum_{y=1}^{H}\sum_{x=1}^{W} Z_{b,c,y,x} \;\in\; \mathbb{R}, \tag{5.3}$$

yielding $\mathrm{GAP}(Z) \in \mathbb{R}^{B\times C}$. i.e., one vector per sample and channel (no learnable parameters), aggregating spatial evidence into a single modality token per stream. The three modality contexts are stacked and processed by multi-head

attention, with the primary context as the query:

$$
\begin{aligned}
C_{\text{stack}} &= \text{Stack}([c_{\text{PR}}, c_{\text{TH}}, c_{\text{UV}}]) \in \mathbb{R}^{B\times 3\times C},\\
Q &= \text{reshape}_h(W_Q\, c_{\text{PR}}) \in \mathbb{R}^{B\times h\times 1\times d_h},\\
K &= \text{reshape}_h(W_K\, C_{\text{stack}}) \in \mathbb{R}^{B\times h\times 3\times d_h},\\
V &= \text{reshape}_h(W_V\, C_{\text{stack}}) \in \mathbb{R}^{B\times h\times 3\times d_h},\\
&\quad \text{(with } h \text{ heads and per-head width } d_h\text{; typically } C = h\, d_h\text{)},\\
A &= \text{softmax}\Big(\tfrac{QK^\top}{\sqrt{d_h}}\Big) \in \mathbb{R}^{B\times h\times 1\times 3},\\
&\quad \text{(softmax over modalities)},\\
U &= AV \in \mathbb{R}^{B\times h\times 1\times d_h},\\
&\quad \text{(batched per-head matmul over the modality dimension)},\\
\hat{c} &= W_O\, \text{merge}_h(U) \;+\; c_{\text{PR}} \;\in\; \mathbb{R}^{B\times C},
\end{aligned}
\tag{5.4}
$$

so attention operates over $M{=}3$ modality tokens (not HW spatial tokens), giving $O(M^2)$ complexity versus $O(N^2)$ with $N{=}\frac{H}{4}\frac{W}{4}$ spatial tokens. The attended context is normalised and broadcast to pre-logit resolution:

$$
F_{\text{gcma}} \;=\; \text{Broadcast}(\text{Norm}(\hat{c})) \;\in\; \mathbb{R}^{B\times C\times \frac{H}{4}\times \frac{W}{4}}. \tag{5.5}
$$

Figure 5.5 illustrates this mechanism.

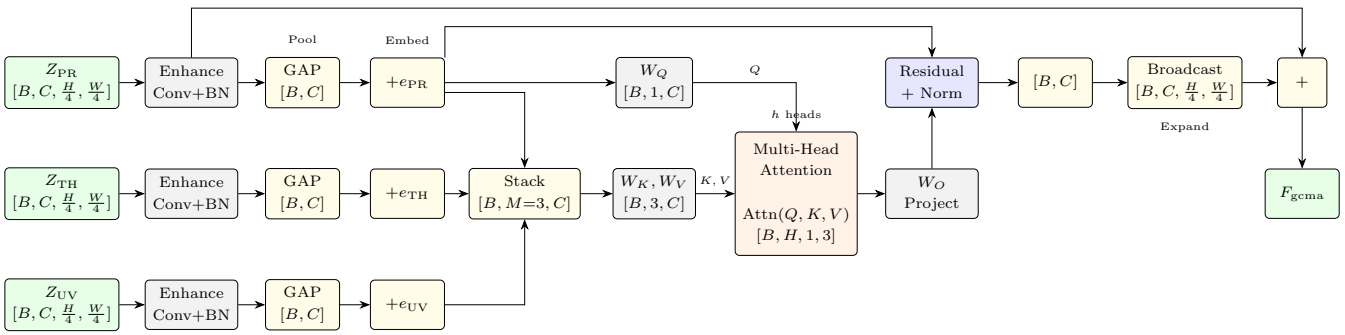


Figure 5.5: Global Context Modality Attention (GCMA). Pre-logits are enhanced and globally pooled to modality contexts $c_i \in \mathbb{R}^{B\times C}$, stacked to $\mathbb{R}^{B\times 3\times C}$. Multi-head attention computes $A = \text{softmax}(QK^\top/\sqrt{d_h})$ where query $Q \in \mathbb{R}^{B\times h\times 1\times d_h}$ from the primary stream attends to keys and values $K, V \in \mathbb{R}^{B\times h\times 3\times d_h}$ from all three modalities, producing attention weights $A \in \mathbb{R}^{B\times h\times 1\times 3}$ over the modality axis. The operation $U = AV$ denotes per-head batched matrix multiplication, computing an attention-weighted sum of V over the modality dimension. Output projection W_O merges the h heads (either $C{=}h\, d_h$ or $W_O : hd_h \to C$), residual connection with c_{PR}, and spatial broadcast yield $F_{\text{gcma}} \in \mathbb{R}^{B\times C\times H/4\times W/4}$.

Sigmoid-Gated Residuals (SIG)

Fine-grained spatial corrections are added via per-pixel sigmoid gates [4] that modulate transformed auxiliary features. For each auxiliary modality, a gate is computed from the concatenation of the GCMA context and the modality's pre-logit:

$$
\begin{aligned}
g_{\text{th}} &= \sigma\big(W_{\text{th}}^{(g)} * [\, F_{\text{gcma}} \parallel Z_{\text{TH}}\,]\big) \in \mathbb{R}^{B\times 1\times \frac{H}{4}\times \frac{W}{4}},\\
g_{\text{uv}} &= \sigma\big(W_{\text{uv}}^{(g)} * [\, F_{\text{gcma}} \parallel Z_{\text{UV}}\,]\big) \in \mathbb{R}^{B\times 1\times \frac{H}{4}\times \frac{W}{4}},\\
r_{\text{th}} &= g_{\text{th}} \odot (W_{\text{th}}^{(t)} * Z_{\text{TH}}), \quad r_{\text{uv}} = g_{\text{uv}} \odot (W_{\text{uv}}^{(t)} * Z_{\text{UV}}),\\
Z_{\text{fused}} &= F_{\text{gcma}} \;+\; r_{\text{th}} \;+\; r_{\text{uv}}.
\end{aligned}
\tag{5.6}
$$

The gate networks employ $r{=}4$ reduction for efficiency. A $1{\times}1$ classifier produces logits from Z_{fused}, which are upsampled to full resolution. Figure 5.6 shows the complete CMAG pipeline.

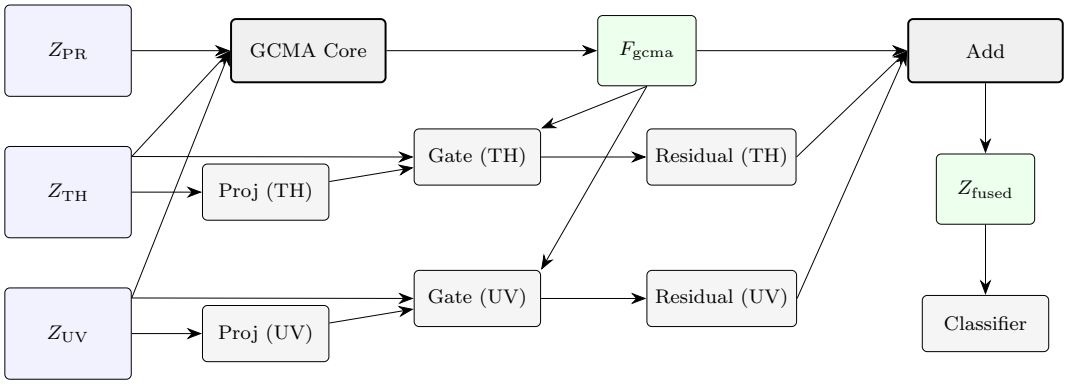


Figure 5.6: CMAG fusion overview. Multi-stage feature pyramids from three modalities are aggregated to pre-logit features ($Z_i \in \mathbb{R}^{B\times C\times H/4\times W/4}$) and projected to $C{=}512$ channels. GCMA pools each modality globally, applies multi-head attention over modality contexts (primary as query), and broadcasts to produce F_{gcma}. Sigmoid-gated residuals r_{TH} and r_{UV} are computed using spatial gates $g_{\text{TH/UV}} \in [0,1]^{B\times 1\times H/4\times W/4}$. Final fusion: $Z_{\text{fused}} = F_{\text{gcma}} + r_{\text{TH}} + r_{\text{UV}}$.

5.5.4 Multi-Head Supervision

We train a fused head alongside three unimodal heads (primary, thermal, UV), all predicting the same semantic classes. Unimodal heads are supervised against modality-specific ground truth, while the fused head uses primary labels. The total objective:

$$\mathcal{L}_{\text{total}} \;=\; \alpha_{\text{pr}}\,\mathcal{L}_{\text{pr}} \;+\; \alpha_{\text{th}}\,\mathcal{L}_{\text{th}} \;+\; \alpha_{\text{uv}}\,\mathcal{L}_{\text{uv}} \;+\; \alpha_{\text{fused}}\,\mathcal{L}_{\text{fused}}, \tag{5.7}$$

with non-negative weights α. This multi-head supervision provides fine-grained gradients to each stream while training the fusion mechanism via $\mathcal{L}_{\text{fused}}$. At inference, auxiliary decoders produce pre-logits for fusion. By default, we output only the fused prediction, with optional per-head outputs for diagnostics at the cost of throughput.

5.5.5 MWPA (Modality-Wise Parallel Attention)

To isolate the contributions of cross-modal attention, we implement MWPA (Modality-Wise Parallel Attention), which applies parallel channel and spatial attention mechanisms rather than cross-modal feature exchange. This baseline enables systematic comparison of attention strategies while maintaining moderate computational efficiency. Given three pre-logit maps

$$Z_{\text{PR}}, Z_{\text{TH}}, Z_{\text{UV}} \in \mathbb{R}^{B\times C\times H/4\times W/4}, \tag{5.8}$$

the method concatenates them along the channel dimension to form $X \in \mathbb{R}^{B\times 3C\times H/4\times W/4}$.

Channel Attention The concatenated features undergo global average pooling followed by a two-layer MLP with reduction ratio $r{=}16$:

$$\mathbf{w}_c = \sigma(\text{Conv}_{1\times1}(\delta(\text{Conv}_{1\times1}(\text{GAP}(X))))), \tag{5.9}$$

where the first convolution reduces from $3C$ to C/r channels, and the second expands back to $3C$ channels. The output $\mathbf{w}_c \in [0,1]^{B\times 3C\times 1\times 1}$ is reshaped to $[B,3,C,1,1]$ to provide per-modality channel attention weights.

Spatial Attention In parallel, the concatenated features are processed through a spatial attention network:

$$\mathbf{m}_s = \sigma(\text{Conv}_{1\times1}(\delta(\text{Conv}_{3\times3}(X)))), \tag{5.10}$$

where the 3×3 convolution (with padding=1) reduces from $3C$ to C/r channels, followed by a 1×1 convolution that produces N=3 spatial attention maps, yielding $\mathbf{m}_s \in [0,1]^{B\times 3\times H/4\times W/4}$.

Modality-Specific Weighting Unlike sequential attention, this method applies both attention types simultaneously to each modality:

$$Z_m^{\text{weighted}} = Z_m \odot \mathbf{w}_c^{(m)} \odot \mathbf{m}_s^{(m)}, \quad m \in \{\text{PR}, \text{TH}, \text{UV}\}, \tag{5.11}$$

where $\mathbf{w}_c^{(m)} \in [0,1]^{B\times C\times 1\times 1}$ and $\mathbf{m}_s^{(m)} \in [0,1]^{B\times 1\times H/4\times W/4}$ are the channel and spatial attention weights for modality m.

Final Fusion The weighted modalities are summed to produce the fused output:

$$Z_{\text{fused}} = \sum_{m\in\{\text{PR,TH,UV}\}} Z_m^{\text{weighted}} \in \mathbb{R}^{B\times C\times H/4\times W/4}. \tag{5.12}$$

This approach enables modality-specific attention learning while maintaining computational efficiency (Figure 5.7).

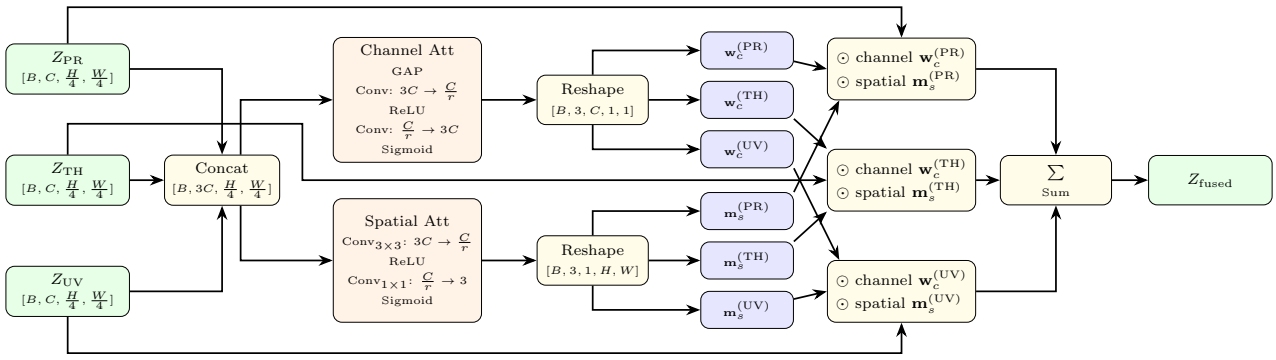


Figure 5.7: **Modality-wise Parallel Attention (MWPA).** From the concatenated pre-logits $[Z_{\text{PR}} \| Z_{\text{TH}} \| Z_{\text{UV}}]$, the module computes for each modality $m \in \{\text{PR}, \text{TH}, \text{UV}\}$:(i) a channel weight $\mathbf{w}_c^{(m)} \in [0,1]^{B\times C\times 1\times 1}$ via GAP→MLP (r=16),and (ii) a spatial mask $\mathbf{m}_s^{(m)} \in [0,1]^{B\times 1\times \frac{H}{4}\times \frac{W}{4}}$ via 3×3/1×1 convolutions with sigmoid. Each modality is recalibrated as $\tilde{Z}_m = (\mathbf{w}_c^{(m)} \odot \mathbf{m}_s^{(m)}) \odot Z_m$ (element-wise with broadcasting), then summed to yield Z_{fused}. *Notation:* $\odot$ denotes element-wise (Hadamard) multiplication with broadcasting; applying channel then spatial (or vice versa) is equivalent.

5.6 Experimental Setup

5.6.1 Comparison Methods and Unified Evaluation Framework

To comprehensively evaluate CMAG, we establish a unified decoder-level fusion framework and adapt five representative methods spanning major fusion paradigms: (i) global channel modulation (MMTM [5]), (ii) recurrent spatial attention (R2AU [6]), (iii) cross-modal attention (GCMA), (iv) sigmoid gating (PL-SIG), and (v) hybrid attention (MWPA). These adaptations enable systematic comparison across diverse fusion paradigms within a consistent architectural framework.

Unified Framework All methods operate within a standardised architecture: MiT-B0 backbone with three independent decoder heads (primary RGB+DIN, thermal, UV) producing C-channel pre-logit features at $1/4$ spatial resolution. Fusion modules integrate these pre-logit features while maintaining consistent spatial dimensions and channel depth across all modalities. A single fusion operation occurs at the pre-logit stage, ensuring a fair comparison across methods. Table 5.3 reports parameter counts and computational complexity (FLOPs) to quantify the overhead introduced by different fusion mechanisms.

Adapted Implementations Methods adapted to our decoder-level fusion framework are designated with a PL- prefix (Pre-Logit) to maintain a clear distinction from their original architectures. Each adaptation faithfully preserves the fundamental fusion mechanism while conforming to our standardised three-modality pre-logit integration scheme, enabling direct performance comparison across heterogeneous fusion strategies.

PL-MMTM adapts the multimodal transfer module [5] from its original two-stream architecture to our trimodal decoder framework via channel-wise squeeze-and-excitation. Given pre-logit features from the three decoder heads,

$$Z_{\mathrm{PR}}, Z_{\mathrm{TH}}, Z_{\mathrm{UV}} \in \mathbb{R}^{B\times C\times H/4\times W/4}, \tag{5.13}$$

the method concatenates them along the channel dimension to form $Z_{\mathrm{concat}} \in \mathbb{R}^{B\times 3C\times H/4\times W/4}$. Global average pooling compresses spatial information into a channel descriptor $\mathbf{s} \in \mathbb{R}^{B\times 3C}$. This descriptor undergoes bottleneck processing through sequential fully-connected layers with reduction ratio r=8:

$$\boldsymbol{\alpha} = \sigma(\mathbf{FC}_2(\delta(\mathbf{FC}_1(\mathbf{s})))), \quad \mathbf{FC}_1 : 3C \to \frac{3C}{r}, \quad \mathbf{FC}_2 : \frac{3C}{r} \to 3C, \tag{5.14}$$

where δ denotes ReLU and σ denotes sigmoid activation. The excitation weights $\boldsymbol{\alpha} \in [0,1]^{B\times 3C}$ are reshaped to $[B, 3C, 1, 1]$ and applied via element-wise multiplication to the concatenated features. The gated features are then split back into individual modalities $Z_{\mathrm{PR}}^{\mathrm{gated}}, Z_{\mathrm{TH}}^{\mathrm{gated}}, Z_{\mathrm{UV}}^{\mathrm{gated}} \in \mathbb{R}^{B\times C\times H/4\times W/4}$ and averaged to produce the final fused representation:

$$Z_{\mathrm{fused}} = \frac{1}{3}(Z_{\mathrm{PR}}^{\mathrm{gated}} + Z_{\mathrm{TH}}^{\mathrm{gated}} + Z_{\mathrm{UV}}^{\mathrm{gated}}) \in \mathbb{R}^{B\times C\times H/4\times W/4}. \tag{5.15}$$

This approach enables each modality to be recalibrated based on the global context of all three inputs before fusion (Figure 5.8).

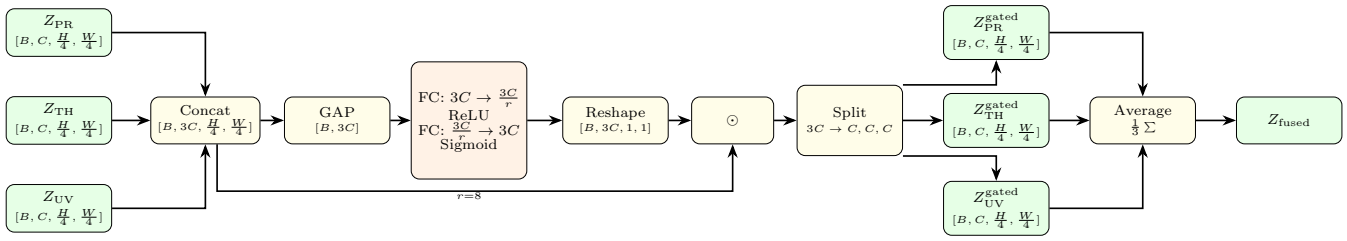


Figure 5.8: **PL-MMTM.** Trimodal fusion through channel-wise squeeze-and-excitation. Features are concatenated, globally pooled, and processed through an MLP bottleneck to generate channel excitation weights. After gating, features are split back to individual modalities and averaged to produce $Z_{\mathrm{fused}} \in \mathbb{R}^{B\times C\times H/4\times W/4}$. *Notation:* $\odot$ denotes element-wise (Hadamard) multiplication with broadcasting.

PL-R2AU adapts recurrent attention gates [6] to decoder-level fusion at the pre-logit stage. Given pre-logit features

$$Z_{\mathrm{PR}}, Z_{\mathrm{TH}}, Z_{\mathrm{UV}} \in \mathbb{R}^{B\times C\times H/4\times W/4}, \tag{5.16}$$

the method employs two attention gates, each using the primary features Z_{PR} as the gating signal g to modulate an auxiliary modality x_m where $m \in \{\mathrm{TH}, \mathrm{UV}\}$. For each gate, both g and x_m undergo separate projections to an intermediate dimension $F_{\mathrm{int}} = C/4$ via 1×1 convolutions with batch normalisation:

$$W_g : \mathbb{R}^C \to \mathbb{R}^{F_{\mathrm{int}}}, \quad W_x : \mathbb{R}^C \to \mathbb{R}^{F_{\mathrm{int}}}. \tag{5.17}$$

The projected features are element-wise summed and processed through ReLU activation, followed by a 1×1 convolution with batch normalisation and sigmoid activation to generate spatial attention masks:

$$\psi_m = \sigma(\mathrm{BN}(\mathrm{Conv}_{1\times 1}(\mathrm{ReLU}(W_g(g) + W_x(x_m))))) \in [0,1]^{B\times 1\times H/4\times W/4}. \tag{5.18}$$

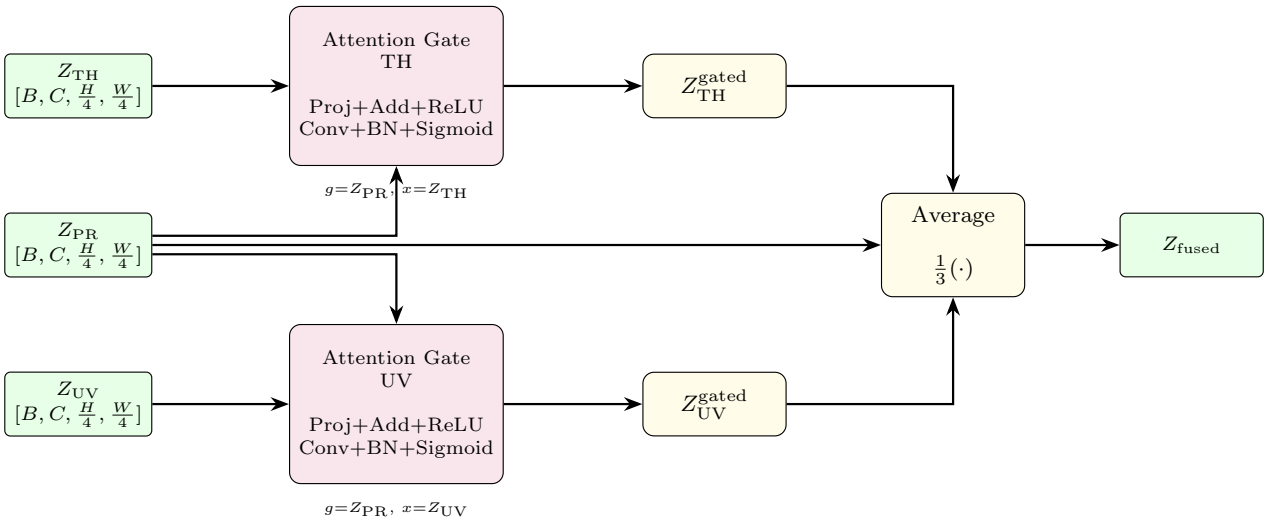


Figure 5.9: **PL-R2AU.** Primary features Z_{PR} serve as gating signals for auxiliary modalities Z_{TH} and Z_{UV} through attention gates. Each gate projects inputs to $F_{\text{int}} = C/4$, computes spatial attention masks via ReLU and sigmoid, then applies element-wise multiplication. The gated auxiliaries are averaged with the primary to yield Z_{fused}.

Each auxiliary modality is gated by its corresponding attention mask: $Z_m^{\text{gated}} = \psi_m \odot Z_m$. The final fused representation averages the primary features with the two gated auxiliaries:

$$Z_{\text{fused}} = \frac{1}{3}(Z_{\text{PR}} + Z_{\text{TH}}^{\text{gated}} + Z_{\text{UV}}^{\text{gated}}) \in \mathbb{R}^{B\times C\times H/4\times W/4}. \tag{5.19}$$

This approach enables selective incorporation of auxiliary information based on primary feature guidance (Figure 5.9).
PL-SIG (pre-logit sigmoid gating). As a stand-alone, attention-free baseline (adapted from GF-Net [4]), we fuse once at the decoder's pre-logit tap (quarter resolution) using the primary pre-logit as the base Z_{PR}. For each auxiliary Z_i, a single-channel gate is predicted from the concatenation $[Z_{\text{PR}} \| Z_i]$ via a two-layer 1×1 MLP (reduction $r{=}4$, ReLU, sigmoid), and the auxiliary is projected on the residual path and masked: $r_i = \psi_i \odot (W_i^{(t)} * Z_i)$. The fused pre-logit is $Z_{\text{fused}} = Z_{\text{PR}} + \sum_i r_i$, which a 1×1 classifier maps to logits before upsampling. Gates are computed from the current base. See Figure 5.10.

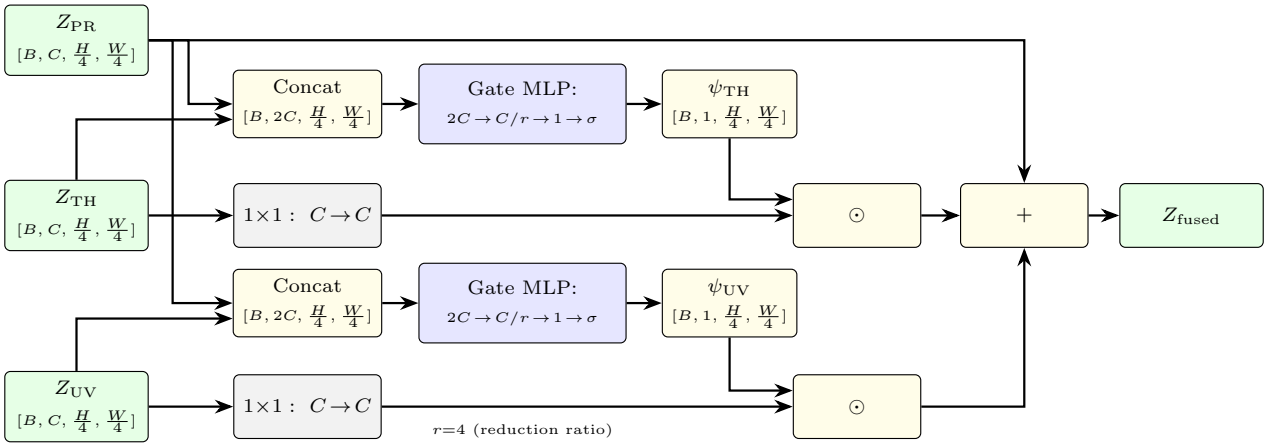


Figure 5.10: **Pre-logit sigmoid gating.** The base is the decoder's primary pre-logit Z_{PR}; each auxiliary Z_i is concatenated with Z_{PR} and passed through a two-layer 1×1 gating MLP (reduction $r{=}4$) to produce a single-channel spatial mask $\psi_i \in [0, 1]$ at quarter resolution. Auxiliaries are projected on the residual path (if needed), masked, and added to form the fused pre-logit Z_{fused}, which is then classified.

Novel Implementations We evaluate two proposed fusion mechanisms: CMAG (Section 5.5.3) and MWPA (Section 5.5.5).

Component Ablations To isolate the individual contributions of CMAG's constituent mechanisms, we evaluate GCMA (Global Context Modality Attention) and sigmoid gating (PL-SIG) as standalone components. Both are described in detail in Section 5.5.3, with GCMA providing the cross-modal attention mechanism and SIG providing per-pixel spatial refinement. These ablations quantify the contribution of each component to CMAG's overall performance.

Training Protocol All methods are trained end-to-end using AdamW optimisation, with a batch size of 6 for 220 epochs. The MiT-B0 backbone employs three independent decoder heads (primary RGB+DIN, thermal, UV), each with per-modality supervision using CEDice loss, manual class weights, and rare-class oversampling. Per-head learning rates are 9×10^{-4} for UV and thermal, and 1.1×10^{-3} for primary, with cosine annealing and head-specific warmup schedules. Multi-head supervision applies loss weights α_{pr}=0.75, α_{th}=0.35, α_{uv}=0.35, and α_{fused}=1.0. Modality dropout (p=0.10) is applied to auxiliary streams during training.

Normalisation swap at evaluation For all the pre-logit fusion modules, LN layers are replaced by GN-16 at test time (channels divisible by 16), carrying over the learnt (γ, β).
Table 5.2 summarises the adapted methods. Source code and trained models will be made publicly available.

Table 5.2: Decoder-level fusion methods adapted for three-modality (RGB+DIN, Thermal, UV) pre-logit integration. Methods with **PL-** prefix are our adaptations preserving core innovations within the unified framework. GCMA and SIG are CMAG's component mechanisms evaluated standalone.

Method	Core Mechanism	Key Innovation	Our Adaptation
PL-MMTM	Global squeeze-excite over concatenated channels	Cross-modal channel transfer via shared global context	Three pre-logits $[Z_{\text{PR}} \| Z_{\text{TH}} \| Z_{\text{UV}}]$ concatenated, SE (r=8), single classifier
PL-R2AU	Recurrent attention gates with primary as gating signal	Spatial attention masks modulate auxiliary contributions	Primary gates thermal/UV via attention gates (F_{int}=256); average fusion of gated features
MWPA	Modality-wise parallel channel and spatial attention	Per-modality attention weights for selective fusion	Concat pre-logits → channel attention (r=16) → spatial attention (3×3 conv) → modality-wise weighting → classifier
PL-SIG	Per-pixel sigmoid gating	Lightweight spatial masks without attention	Base feature + gated residuals from thermal/UV; r=4 reduction
GCMA	Global context cross-modal attention	Modality-level attention ($O(M^2)$ vs $O(N^2)$)	Pool → Stack ($B\times3\times C$) → Multi-head attn (h=4) → Broadcast → classifier
CMAG	GCMA + SIG	Global attention for correspondence + fine-grained spatial gating	GCMA context + sigmoid-gated residuals; unified training

5.6.2 Robustness Evaluation

Robustness to sensor degradation, geometric misalignment, and incomplete multimodal input is critical for deployment applications. We systematically evaluate architectural resilience across three perturbation categories under controlled degradation, spanning three RGB lighting conditions (underexposed/RGB1, optimal/RGB3, overexposed/RGB5).

Modality Dropout Analysis To simulate sensor failure and assess graceful degradation, we systematically ablate individual modalities and measure the resultant impact on fusion performance. Each modality (RGB, DIN, thermal, UV) is independently removed at inference while maintaining all other inputs, testing each method's capacity to preserve functionality under incomplete multimodal input. This establishes the relative importance and contribution of each sensor stream within the fusion framework, quantifying how critically the architecture depends on each modality.

Noise Resilience Testing We evaluate performance degradation under four noise types—Gaussian, salt-and-pepper, speckle, and generic additive noise—applied independently to individual modalities (RGB, DIN, thermal, and UV). Noise intensity is varied across fourteen levels, with fine-grained sampling at low intensities (0.1, 0.2, 0.3, 0.4, 0.5) transitioning to coarser increments at higher intensities (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0), capturing both subtle corruption and severe degradation regimes. Figure 5.11 illustrates representative noise perturbations at three intensity levels across all modalities, demonstrating the range of corruption severity evaluated. This design yields 224 noise configurations (4 types, $\times$ 14 intensities, $\times$ 4 modalities), evaluated across three lighting conditions, resulting in 672 scenarios per method. With six methods under comparison (CMAG, GCMA, MWPA, PL-MMTM, PL-R2AU, PL-SIG), this totals 4,032 noise robustness evaluations, providing a comprehensive characterisation of architectural sensitivity to sensor corruption.

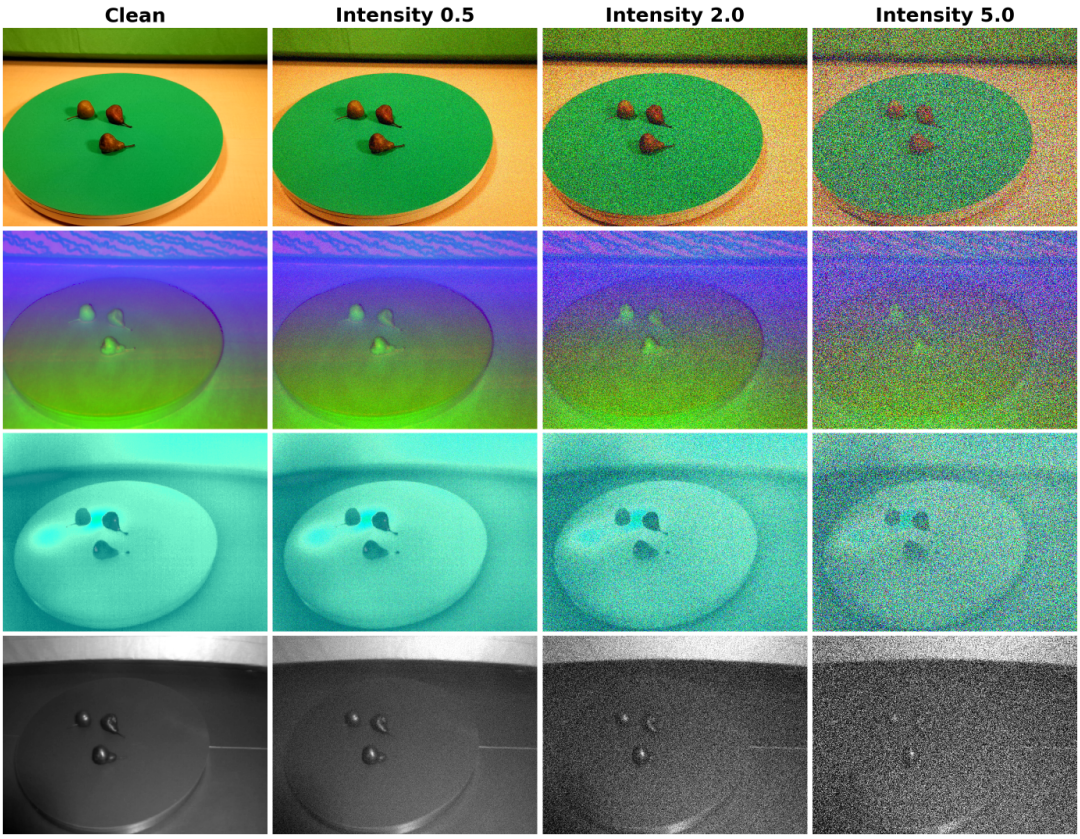


Figure 5.11: Representative noise perturbations applied during robustness evaluation. Rows show different modalities (RGB, DIN, thermal, UV); columns show clean input and three noise intensity levels (0.5, 2.0, 5.0) for Gaussian noise. At intensity 0.5, corruption is subtle; at 2.0, significant degradation is visible; at 5.0, severe corruption challenges recognition. Similar patterns apply to salt-and-pepper, speckle, and uniform noise types (not shown). All RGB examples are from the RGB3 (optimal lighting) test set.

Spatial Misalignment Testing We assess robustness to geometric misregistration by applying controlled spatial shifts to thermal and UV modalities, simulating realistic sensor calibration drift or mechanical misalignment. Each auxiliary modality is independently displaced by 20 and 40 pixels in the four cardinal directions (up, down, left, right), while the primary RGB-DIN stream remains stationary as the reference coordinate frame. Figure 5.12 visualises the effect of spatial misalignment on thermal and UV inputs, illustrating how features become spatially inconsistent with the primary stream. This produces 16 shift configurations (2 distances $\times$ 4 directions $\times$ 2 modalities), evaluated across three lighting conditions, yielding 48 misalignment scenarios per method. Across six methods, this generates 288 spatial robustness evaluations. These test-time shifts evaluate each method's intrinsic capacity to maintain performance under geometric inconsistency without calibration or retraining.

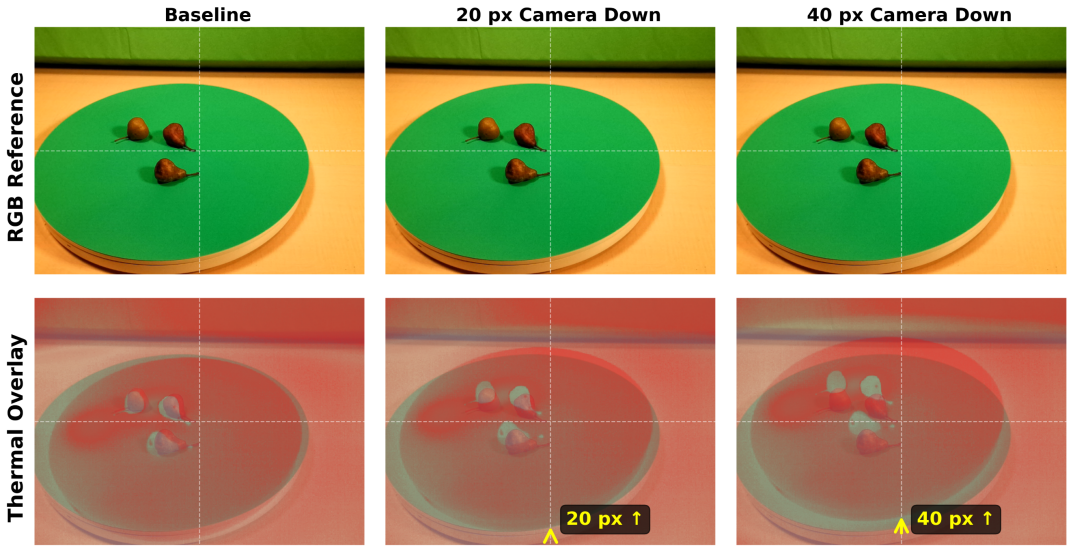


Figure 5.12: Spatial misalignment examples for thermal modality. Left: aligned baseline showing RGB and thermal overlay with default trained spatial registration. Middle: 20-pixel upward shift of thermal modality, creating moderate misregistration. Right: 40-pixel upward shift of thermal modality, demonstrating severe geometric inconsistency. Coloured overlays highlight spatial discrepancies between primary and auxiliary features. The RGB image remains fixed as the reference coordinate frame throughout all shift scenarios.

Impact Metrics For a given RGB lighting condition r and perturbation scenario a, we quantify performance degradation relative to the unperturbed baseline ($\mathrm{mIoU}_r^{\mathrm{Full}}$) using two complementary metrics:

$$\mathrm{Impact}_{\%}(r,a) = 100\,\frac{\mathrm{mIoU}_r^{\mathrm{Full}} - \mathrm{mIoU}(r,a)}{\mathrm{mIoU}_r^{\mathrm{Full}}}, \tag{5.20}$$

$$\mathrm{Impact}_{\mathrm{pp}}(r,a) = \mathrm{mIoU}_r^{\mathrm{Full}} - \mathrm{mIoU}(r,a), \tag{5.21}$$

where $\mathrm{Impact}_{\%}$ quantifies relative degradation (normalised by baseline performance) and $\mathrm{Impact}_{\mathrm{pp}}$ measures absolute loss in percentage points. We prioritise $\mathrm{Impact}_{\mathrm{pp}}$ for cross-method comparisons, as it avoids the baseline bias inherent in relative metrics; methods with lower baseline accuracy may appear artificially robust when evaluated via percentage degradation. Per-scenario results are aggregated as mean $\pm$ standard deviation across the full test split, with lighting-specific statistics reported separately to isolate illumination-dependent impacts.

5.7 Results and Discussion

We evaluate six decoder-level fusion architectures across three lighting conditions. Systematic robustness analysis quantifies performance under modality dropout, spatial misalignment, and noise corruption. Results are organised by baseline performance, ablation studies, and comparative discussions of fusion strategies.

5.7.1 Evaluation Metrics

We report five standard segmentation metrics for comprehensive evaluation. Our primary metric, mean Intersection over Union (mIoU), provides class-balanced accuracy by averaging IoU across all classes. Frequency-Weighted IoU (FIoU) emphasises prevalent classes by weighting performance by pixel frequency. Pixel-level metrics include mean pixel accuracy (MPA), which computes per-class recall without penalising false positives, and pixel accuracy (PA), which measures overall pixel-level correctness. For distribution-free comparison, we report Mean Rank [32], where methods are ranked by IoU within each class and ranks are averaged across classes (a lower value indicates superior performance).

5.7.2 Overall Performance

Table 5.3 presents comprehensive baseline performance across all six fusion architectures under three lighting conditions. CMAG achieves the highest average mIoU of 84.18% across lighting conditions, validating its hybrid fusion strategy, which combines global context attention with fine-grained spatial gating. Under optimal lighting (RGB3), both CMAG (87.61%) and GCMA (86.72%) achieve strong performance, with a modest 0.89 percentage point difference. However, CMAG demonstrates superior robustness under challenging illumination; while GCMA degrades substantially (RGB1: 80.49%, RGB5: 78.03%), CMAG maintains more stable performance (RGB1: 82.54%, RGB5: 82.38%), providing a 3.2 percentage point advantage that is 3.6-fold larger than the optimal-lighting gap.
The remaining architectures achieve competitive baseline accuracy, ranging from 81.43% (PL-MMTM) to 83.14% (PL-R2AU) average mIoU, with reduced computational overhead (19M parameters vs. 22M for CMAG/GCMA). PL-R2AU demonstrates consistent performance across lighting conditions (average: 83.14%), while MWPA achieves an average mIoU.of 82.06%. PL-MMTM shows moderate baseline accuracy (average: 81.43%) but demonstrates superior robustness under perturbations, as detailed in Section 5.7.7. The lightweight sigmoid gating baseline (PL-SIG) achieves 82.29% average mIoU with balanced performance across conditions.
Mean rank analysis reveals a consistent performance ordering. Under optimal lighting (RGB3), CMAG achieves the best mean rank (3.12), followed by GCMA (4.56), PL-R2AU (5.55), and MWPA (5.88). Under suboptimal lighting, CMAG maintains strong rankings (RGB1: 10.28, RGB5: 9.05), while GCMA exhibits greater rank sensitivity to lighting degradation (RGB1: 12.59, RGB5: 12.12). All architectures sustain real-time throughput (31–34 FPS); detailed computational costs and throughput are reported in Section 5.7.6.
Frequency-weighted IoU (FIoU) is near-ceiling across all methods (typically 99.3–99.6%), reflecting class imbalance dominated by background and large, frequent items. The resulting mIoU–FIoU gap (about 12–16 pp across methods) indicates that errors concentrate in rarer or visually ambiguous classes; we analyse these in Section 5.7.5.

Table 5.3: Network comparison across lighting conditions and fusion architectures. Overall metrics (mIoU, FIoU, MPA, PA) and mean rank(1–18, lower is better) scores are reported across the three light settings of the six fusion methods. Best value per RGB configuration in bold. RGB1: underexposed; RGB3: optimal; RGB5: overexposed lighting.

Metric	RGB1 CMAG	RGB3 CMAG	RGB5 CMAG	RGB1 GCMA	RGB3 GCMA	RGB5 GCMA	RGB1 PL-R2AU	RGB3 PL-R2AU	RGB5 PL-R2AU	RGB1 MWPA	RGB3 MWPA	RGB5 MWPA	RGB1 PL-SIG	RGB3 PL-SIG	RGB5 PL-SIG	RGB1 PL-MMTM	RGB3 PL-MMTM	RGB5 PL-MMTM
Mean Rank	**10.3**	**3.1**	**9**	12.6	4.6	12.1	11.8	5.6	10.5	13.5	5.9	9.9	10.9	6.2	10.9	12.6	8.5	13.1
mIoU (%)	**82.54**	**87.61**	**82.38**	80.49	86.72	78.03	81.39	86.02	82.02	78.99	85.71	81.47	81.27	85.31	80.29	80.76	84.09	79.45
FIoU (%)	**99.37**	**99.55**	**99.37**	99.29	99.53	99.28	99.30	99.51	99.35	99.26	99.50	99.36	99.33	99.48	99.32	99.27	99.46	99.28
MPA (%)	**89.21**	**92.58**	**88.98**	88.00	91.87	85.18	88.13	91.86	88.21	86.21	91.70	87.30	88.75	91.45	87.04	87.76	90.17	86.54
PA (%)	**99.66**	**99.76**	**99.66**	99.61	99.75	99.61	99.62	99.74	99.65	99.60	99.73	99.65	99.64	99.72	99.63	99.61	99.70	99.61
FPS	31	31	31	31	31	31	34	34	34	32	32	32	34	34	34	34	34	34
Params (M)	22	22	22	22	22	22	19	19	19	19	19	19	19	19	19	19	19	19
GFLOPs	91.7	91.7	91.7	89.2	89.2	89.2	79.0	79.0	79.0	84.0	84.0	84.0	84.0	84.0	84.0	74.0	74.0	74.0

5.7.3 Normalisation Strategy: LayerNorm vs GroupNorm Trade-offs

Table 5.4 compares the impact of normalisation choice on accuracy and throughput across all six fusion architectures. LayerNorm (LN) during training consistently yields the highest mIoU across methods, establishing it as the preferred normalisation for learning decoder-level fusion. However, LN incurs substantial computational cost at inference: on $C{=}512$ feature maps at $\frac{H}{4}\times\frac{W}{4} = 120\times160$ resolution, LN achieves only 13.5–14.1 FPS across methods. Switching to GroupNorm (GN) at inference while retaining LN-trained weights provides a favourable accuracy-throughput trade-off. GN-16 maintains accuracy within 0.2 percentage points of LN for most methods (CMAG: 87.61% $\rightarrow$ 87.45%, GCMA: 86.84% $\rightarrow$ 86.72%), while improving throughput by approximately 2.4 times. Finer grouping (GN-8) offers minimal accuracy improvement over GN-16 at reduced throughput, while coarser grouping (GN-32) marginally increases speed but exhibits slight accuracy degradation for some methods (MWPA: 85.71% $\rightarrow$ 85.35%, PL-SIG: 85.31% $\rightarrow$ 85.24%).

Training with GroupNorm directly (rather than LN) was briefly explored but yielded inferior performance, suggesting that global normalisation statistics during learning are important for cross-modal attention mechanisms. The asymmetric LN-train/GN-16 inference scheme, therefore, represents the optimal configuration, balancing accuracy with real-time inference requirements.

Table 5.4: Normalisation method comparison under optimal lighting (RGB3, 220 epochs). LayerNorm (LN) used during training for all configurations; GroupNorm (GN) variants applied at inference only. Best mIoU per method in bold. All methods show 2.3–2.4× throughput improvement with GN-16 inference while maintaining accuracy within 0.2pp of LN.

Method	**LN (Inference)**				**GN-8**				**GN-16**				**GN-32**			
	mIoU	FIoU	MPA	FPS	mIoU	FIoU	MPA	FPS	mIoU	FIoU	MPA	FPS	mIoU	FIoU	MPA	FPS
CMAG	**87.61**	99.55	92.58	13.51	87.43	99.54	92.69	29.90	87.45	99.54	92.52	31.09	87.42	99.53	92.38	31.32
GCMA	**86.84**	99.54	91.89	13.78	86.77	99.54	91.83	30.93	86.72	99.53	91.87	32.06	86.68	99.53	92.02	32.64
MWPA	85.67	99.51	90.71	13.92	**85.72**	99.51	91.16	31.80	85.71	99.50	91.70	33.01	85.35	99.47	92.21	33.32
PL-SIG	**85.59**	99.50	91.03	13.97	85.56	99.50	91.23	31.98	85.31	99.48	91.45	33.20	85.24	99.46	91.78	33.83
PL-MMTM	84.05	99.46	89.51	14.07	84.07	99.46	89.80	32.54	**84.09**	99.46	90.17	33.80	83.78	99.41	91.01	34.49
PL-R2AU	**86.07**	99.53	91.38	13.98	86.00	99.52	91.65	32.26	86.02	99.51	91.86	33.49	85.90	99.50	92.17	34.19

5.7.4 Lighting Condition Sensitivity

RGB illumination quality substantially affects segmentation performance across all decoder-level fusion architectures. We evaluate three illumination settings (RGB1: underexposed, RGB3: optimal, RGB5: overexposed) with results in Tables 5.3 and 5.12.

Performance Across Lighting Conditions Under optimal illumination (RGB3), methods achieve peak performance with narrow differentiation: CMAG (87.61% mIoU), GCMA (86.72%), PL-R2AU (86.02%), MWPA (85.71%), PL-SIG (85.31%), and PL-MMTM (84.09%) — a spread of only 3.52 pp. Underexposure (RGB1) shows a similar spread of 3.55 pp, with CMAG maintaining 82.54% (5.07 pp loss) and MWPA degrading to 78.99% (6.72 pp loss). Overexposure (RGB5) expands the spread to 4.35 pp, with GCMA suffering the largest degradation to 78.03% (8.69 pp loss). Averaging across methods, RGB3→RGB5 degradation (5.30 pp) slightly exceeds RGB3→RGB1 (5.00 pp), reflecting irreversible saturation-induced information loss that auxiliary modalities cannot fully recover.
Mean rank analysis reveals differences in architectural stability: CMAG maintains consistent rankings (RGB3: 3.12, RGB1: 10.28, RGB5: 9.05), whereas GCMA exhibits high volatility (RGB3: 4.56, RGB1: 12.59, RGB5: 12.12), demonstrating a strong dependence on RGB quality.

Surface-Dependent Vulnerability Class-level analysis reveals that surface properties dominate lighting robustness, far exceeding architectural differences. Reflective objects distinguished primarily by colour and pattern rather than geometric shape exhibit severe overexposure sensitivity: Apple degrades 27.50 pp (96.72%→69.22% at RGB5) for CMAG, as saturation obliterates the discriminative colour and patterns. In contrast, geometrically similar but less reflective Apple Green loses only 1.79 pp (95.80%→94.01%). Texture-rich objects maintain robust performance across lighting extremes: Grapes Blue achieves stable segmentation (RGB1: 94.24%, RGB3: 94.38%, RGB5: 95.92%), as geometric texture features survive both underexposure and saturation, where colour and specularity fail.
Sophisticated fusion mechanisms provide measurable advantages when auxiliary modalities are essential for discrimination. Under optimal lighting (RGB3), Carrot and Carrot Fake are visually nearly identical in RGB-DIN; yet, thermal signatures differ markedly (plastic vs. organic emissivity). Here, the architectural capacity to leverage thermal cues determines performance: CMAG achieves 90.05% on Carrot and 79.74% on Carrot Fake, while GCMA achieves 84.89% and 59.38% respectively, a 20.36 percentage point gap in replica discrimination performance. This demonstrates that CMAG's hybrid attention-gating mechanism more effectively weights thermal features when visual appearance alone is insufficient. Under underexposure (RGB1), where RGB colour and intensity cues are severely degraded while visual similarity between real and replica persists, PL-SIG excels on Carrot Fake (80.90%) through effective thermal gating, whereas PL-MMTM struggles (64.97%). This pattern highlights that the architectural capacity to modulate auxiliary contributions becomes critical when primary RGB features degrade, forcing greater reliance on thermal discrimination.

However, no architecture overcomes severe material-specific failures when primary modality features collapse; reflective Apple under overexposure saturation (27.50 pp loss) and Onion Red similarly degraded by saturation (11.62 pp loss) demonstrate fundamental limits where auxiliary modalities cannot compensate for irreversible information loss in the primary stream. Lighting robustness is therefore jointly determined by the architectural capacity to leverage auxiliary modalities when discriminative, the material properties governing primary feature preservation under lighting extremes, and the information-theoretic limits when primary features are irreversibly lost to sensor saturation or severe underexposure.

5.7.5 Challenging Classes and Modality-Specific Contributions

Background and container objects achieve near-ceiling accuracy across methods (Background: 99.77-99.86%, Bowl: 89.35-93.22%), whereas several categories expose persistent multimodal fusion challenges.

Partially Decayed Fruit Partially rotten items (Apple Green Bad, Lemon Bad, Mandarin Bad) span a wide range of difficulty (49.74–94.26% IoU envelope across methods and lighting). Under optimal lighting (RGB3), Apple Green Bad exceeds 90% IoU for all six methods (range: 90.60-94.26%). Yet, performance separates dramatically under suboptimal conditions: at RGB1, the range widens to 60.01-72.17%, with PL-SIG being the most stable (72.17%/92.93%/73.07% at RGB1/3/5) and MWPA showing the most extensive spread (60.01%/94.26%/75.39%). For Mandarin Bad, RGB5 proves particularly fragile (GCMA: 23.14%), highlighting severe exposure sensitivity on decayed surfaces. The thermal modality provides critical complementary information here, as decay alters surface temperature and emissivity patterns that remain discriminative when RGB features degrade. Networks with explicit gating mechanisms (CMAG, PL-SIG) maintain superior cross-lighting stability, suggesting effective learnt thermal utilisation.

Synthetic Replica Objects Replica plastics (Carrot Fake, Apple Fake, Lemon Fake) prove consistently challenging, with class envelopes of 59.38-80.90% (Carrot Fake), 76.54-94.34% (Apple Fake), and 41.13-74.34% (Lemon Fake). On Carrot Fake, PL-SIG achieves 80.90% at RGB1, while GCMA struggles at RGB3 (59.38%), indicating that challenging material discrimination benefits from methods that modulate auxiliary contributions effectively. However, the optimal architecture varies with illumination. The difficulty stems from the near-identical visual appearance of organic counterparts. Plastic exhibits distinctly different thermal emission (lower emissivity, faster thermal equilibration) and UV fluorescence, making thermal and UV modalities essential for this discrimination task.

Temperature-Discriminable Objects For thermally separable classes, cross-method spreads are small. Cup Hot spans 88.62-94.75% overall, with per-lighting ranges of RGB1: 88.62-94.46% (5.8 pp), RGB3: 92.59-94.98% (2.4 pp), RGB5: 91.54-93.88% (2.3 pp). Cup Cold spans 82.90-95.08% with RGB1: 82.90-92.12% (9.2 pp), RGB3: 92.47-95.06% (2.6 pp), RGB5: 90.94-95.08% (4.1 pp). Under optimal or overexposed lighting, method variation is $\lesssim$2.6 pp, while underexposure increases the spread to 5.8 to 9.2 pp. This pattern validates that thermal provides unambiguous discriminative features when temperature differences are pronounced.

Implications for Fusion Design Difficult classes exhibit wide intra-class performance ranges: Carrot Fake varies by $\sim$22 pp and Apple Green Bad by $\sim$34 pp across methods and lighting, compared with $\leq$0.8 pp for Mirror (98.19-98.96% range). Methods with gating or attention-based fusion (CMAG, GCMA, PL-SIG) rank better in challenging categories under favourable lighting (RGB3 mean ranks: CMAG 3.12 vs. PL-MMTM 8.50, yet no single architecture dominates across all lighting regimes and classes. The 6-9× greater performance variation on ambiguous categories compared to simple objects demonstrates that the architectural capacity to selectively leverage auxiliary modalities critically determines performance when RGB cues are insufficient. However, exposure-induced failures at RGB5 can overwhelm auxiliary compensation for certain materials, establishing fundamental limits of decoder-level fusion regardless of architectural sophistication.

5.7.6 Computational Efficiency and Real-Time Performance

All six decoder-level fusion architectures sustain real-time throughput on a single NVIDIA RTX 3090 GPU at 640×480 resolution (Table 5.3). Parameter counts cluster in two tiers: 22M for CMAG and GCMA (attention-based

methods), and 19M for MWPA, PL-R2AU, PL-SIG, and PL-MMTM (lighter fusion mechanisms). Computational cost spans 74.0 to 91.7 GFLOPs, with CMAG (91.7 GFLOPs) and GCMA (89.2 GFLOPs) representing the upper bound, MWPA (84.0 GFLOPs), PL-R2AU (79.0 GFLOPs), and PL-SIG (84.0 GFLOPs) occupying the middle range, and PL-MMTM (74.0 GFLOPs) offering the most efficient configuration. Despite these variations, measured throughput remains tightly bounded at 31-34 FPS across all methods, demonstrating that decoder-level fusion introduces minimal overhead beyond the shared encoder-decoder backbone.

The accuracy-efficiency trade-off favours adaptive fusion mechanisms. CMAG achieves the highest average mIoU of 84.18% (mean across RGB1/3/5) at 31 FPS with 91.7 GFLOPs. In comparison, the most efficient method, PL-MMTM, operates at 34 FPS with 74.0 GFLOPs while achieving an average mIoU of 81.43%. Relative to PL-MMTM, CMAG incurs a 24% computational overhead (91.7 vs. 74.0 GFLOPs) and an 8.8% throughput reduction in exchange for a 2.75 pp mIoU improvement. Other methods occupy intermediate positions: PL-R2AU (79.0 GFLOPs, 34 FPS, 83.14% mIoU) and MWPA (84.0 GFLOPs, 32 FPS, 82.06% mIoU) offer balanced alternatives, while GCMA (89.2 GFLOPs, 31 FPS, 81.75% mIoU) demonstrates that computational cost alone does not guarantee superior accuracy; architectural design and lighting robustness are critically important.

All configurations maintain throughput exceeding 30 FPS at VGA resolution, meeting real-time requirements for online robotic tasks. The modest computational differences between methods (17.7 GFLOPs range, 3 FPS variation) relative to substantial accuracy variations (2.75 pp between best and worst) underscore that the decoder-level fusion strategy, rather than raw computational capacity, determines segmentation quality in challenging multimodal scenarios.

5.7.7 Modality Importance: Ablation Studies

We quantify each modality's contribution and alignment sensitivity through two ablation families: (i) drop ablations (complete removal of a modality at inference) and (ii) spatial shift ablations. Unless stated otherwise, results are averaged across all six networks and three lighting conditions. Comprehensive class-wise robustness visualisations for all decoder-level architectures are provided in Appendix 5.B.3.

Drop Ablation Results

Complete modality removal establishes a consistent importance hierarchy: RGB > DIN > T24 > U8. Figure 5.13(left) presents the consolidated results, with mean absolute mIoU losses averaged across networks and lighting conditions:

- **RGB drop:** 59.50 pp loss (72.05% relative degradation)
- **DIN drop:** 49.61 pp loss (60.53% relative degradation)
- **T24 drop:** 24.62 pp loss (29.79% relative degradation)
- **U8 drop:** 16.82 pp loss (20.36% relative degradation)

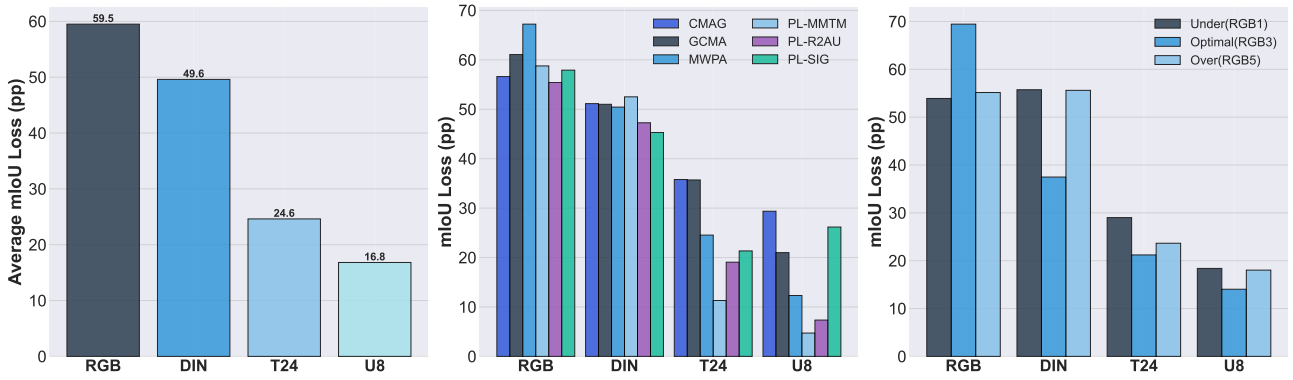


Figure 5.13: Overview of drop ablation impact. (left) Overall modality importance averaged across all networks and lighting conditions, ranked from most to least critical. (centre) Network comparison showing impact for all six fusion networks across the four modalities, revealing architecture-specific dependencies. (right) RGB variant comparison showing how lighting conditions (underexposed, optimal, overexposed) affect modality importance when averaged across networks.

The dominance of RGB and DIN reflects their role as the primary feature extractors for geometric structure and appearance, while thermal and UV provide complementary discriminative cues for challenging material classes (Section 5.7.5). Table 5.5 presents per-network residual mIoU after each modality is dropped, revealing substantial architectural variation: CMAG and GCMA exhibit the highest sensitivity (mean losses of 43.22 pp and 42.19 pp, respectively), while PL-MMTM and PL-R2AU demonstrate superior resilience (31.82 pp and 32.27 pp).
One illustrative example demonstrates the magnitude of primary modality dependence: CMAG under optimal lighting (RGB3) degrades from 87.61% mIoU baseline to 17.49% when RGB is removed (a 70.12 pp loss), confirming the critical role of the primary visual stream.

Table 5.5: Drop ablation results averaged over RGB scenarios: residual mIoU (%) after removing each modality. Networks ordered by robustness (left to right: least robust to most robust). Lower values indicate greater modality dependence. The bottom row shows the mean loss.

Dropped	**RGB**	**CMAG**	**GCMA**	**MWPA**	**PL-SIG**	**PL-R2AU**	**PL-MMTM**
RGB	RGB1	34.51	30.09	11.81	37.19	31.09	19.75
	RGB3	17.49	11.73	11.81	11.91	19.41	21.80
	RGB5	30.70	20.15	23.38	25.04	32.65	21.80
DIN	RGB1	18.54	24.01	27.63	27.89	33.40	22.05
	RGB3	57.86	45.37	42.10	56.16	52.26	32.17
	RGB5	22.73	22.86	27.63	27.89	21.98	27.91
T24	RGB1	36.30	47.97	62.48	55.85	48.07	63.23
	RGB3	61.99	31.51	65.74	72.15	77.61	74.60
	RGB5	46.89	58.68	46.83	55.85	66.57	67.92
U8	RGB1	55.48	56.48	70.57	50.03	74.44	70.61
	RGB3	53.68	70.79	70.57	69.31	82.05	80.21
	RGB5	55.27	55.00	70.57	50.03	70.89	74.70
Mean loss (pp)		43.22	42.19	38.63	37.68	32.27	31.82

Spatial Shift Ablations

We assess robustness to sensor misalignment by applying controlled spatial shifts to thermal and UV images (20 px/40 px), while maintaining RGB-DIN as the reference coordinate frame, with results aggregated across shift configurations, networks, and lighting conditions.

Overall Impact Spatial misalignment induces substantially lower degradation than complete modality removal (mean: 4.20 pp vs. 37.64 pp for drops), confirming that decoder-level fusion exhibits intrinsic tolerance to moderate sensor misregistration. Table 5.6 presents detailed per-network degradation under thermal and UV shifts. Degradation scales non-linearly with magnitude: 20 px shifts cause a 2.61 pp mean loss, while 40 px shifts induce a 5.68 pp loss, a 2.18$\times$ increase. This super-linear relationship reflects that larger misalignments increasingly violate the spatial correspondence assumptions implicit in learnt fusion weights. The proportion of fusion failures (where multimodal performance falls below single-modality baselines) increases 3.6-fold from 8.2% at 20 px to 29.3% at 40 px, establishing a critical misalignment threshold beyond which auxiliary information degrades rather than enhances segmentation accuracy.

Modality-Specific Sensitivity As shown in Table 5.6, thermal exhibits consistently greater sensitivity to misalignment than UV across all networks (mean degradation: thermal 3.57 pp at 20 px and 7.67 pp at 40 px vs. UV 1.66 pp at 20 px and 3.68 pp at 40 px). This differential reflects thermal's greater contribution to fusion performance, as demonstrated by drop ablations (Section 5.7.7); removing thermal causes a 24.62 pp mean loss compared to UV's 16.82 pp loss. Modalities with larger fusion contributions exhibit proportionally greater sensitivity to spatial misalignment, as misregistration directly degrades the discriminative features that the network has learnt to rely upon.

Directional Asymmetry Thermal demonstrates a mild vertical bias, with upward shifts proving most damaging (20 px: 3.89 pp, 40 px: 7.45 pp), while UV exhibits near-isotropic behaviour (directional spread <0.6 pp at each magnitude), as shown in Figure 5.14. This divergence reflects the different object categories that contribute discriminative

Table 5.6: Performance degradation (pp) under spatial misalignment of auxiliary modalities. Values show mean mIoU loss ± standard deviation when thermal or UV streams are shifted relative to the aligned RGB-DIN primary. All networks demonstrate moderate degradation at 20 px shifts (1.07–3.95 pp), which approximately doubles at 40 px shifts (2.71–8.40 pp), confirming the inherent tolerance of decoder-level fusion to geometric misalignment.

Network	**Thermal Shift**		**UV Shift**		**Mean Degradation**	
	20 px	40 px	20 px	40 px	20 px	40 px
CMAG	3.50±0.77	7.99±1.57	1.61±0.89	4.73±1.76	2.56	6.36
GCMA	3.05±0.86	6.92±1.54	1.07±0.73	3.06±1.64	2.06	4.99
MWPA	3.61±0.94	7.86±1.74	1.43±0.78	3.05±1.31	2.52	5.46
PL-MMTM	3.44±0.87	7.35±1.64	1.22±0.45	2.71±0.97	2.33	5.03
PL-R2AU	3.86±0.63	8.40±1.72	1.60±0.85	3.82±1.17	2.73	6.11
PL-SIG	3.95±0.68	7.49±2.92	3.00±1.40	4.72±1.24	3.48	6.11
Average	3.57±0.33	7.67±0.51	1.66±0.71	3.68±0.89	2.61	5.68

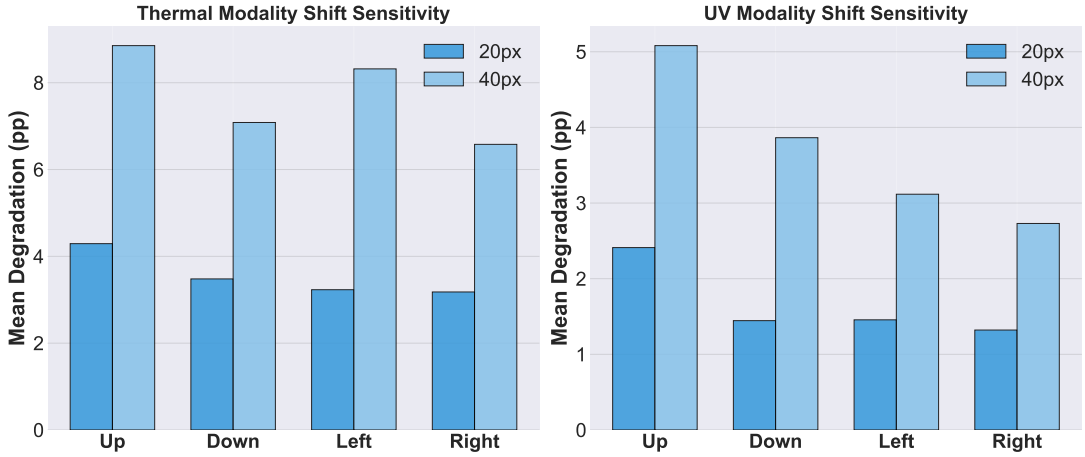


Figure 5.14: Spatial shift sensitivity by magnitude (20 px/40 px) and direction (up, down, left, right) for thermal and UV modalities, averaged across all networks and lighting conditions. Thermal exhibits mild vertical bias (upward shifts most damaging: 7.45 pp at 40 px) and consistently 1.63× greater sensitivity than UV across all scenarios. Degradation scales super-linearly with offset magnitude, doubling the shift produces 1.93× greater loss, indicating that geometric inconsistency tolerance degrades non-linearly beyond modest misalignments.

cues for each modality. Thermal-discriminable objects (temperature-dependent items) and UV-discriminable objects (material-dependent items) occupy different spatial distributions, sizes, and positions within the scene. The thermal vertical bias likely reflects that temperature-based objects in the dataset exhibit systematic vertical positioning patterns, while UV-critical discrimination tasks (replica vs. organic material) occur across more spatially diverse object orientations. This pattern persists across all networks and lighting conditions, confirming that it originates from the dataset object distribution rather than architectural characteristics.

Architectural Robustness Figure 5.15 presents comprehensive per-network robustness across all shift scenarios. The left heatmap shows mean fusion degradation (pp) for each network-scenario combination, while the right heatmap displays the percentage of scenarios where fusion remains beneficial (shifted multimodal mIoU exceeds single-modality baseline). A score of 100% indicates that fusion provides a positive benefit in all evaluated scenarios (across lighting conditions and directions) for that modality-magnitude combination, while lower percentages reveal cases where spatial misalignment causes fusion to underperform relative to single-modality baselines.

Network robustness varies substantially: PL-SIG achieves 94% positive scenarios across all shifts, followed by GCMA (89%), CMAG (83%), PL-R2AU (80%), MWPA (72%), and PL-MMTM (70%). Notably, this ranking differs from baseline accuracy ordering (Table 5.3), demonstrating that peak performance and shift robustness are partially orthogonal properties. Simple gating mechanisms (PL-SIG) prove most tolerant to misalignment, while sophisticated fusion (CMAG) trades shift robustness for higher clean-data accuracy. The heatmap reveals that UV

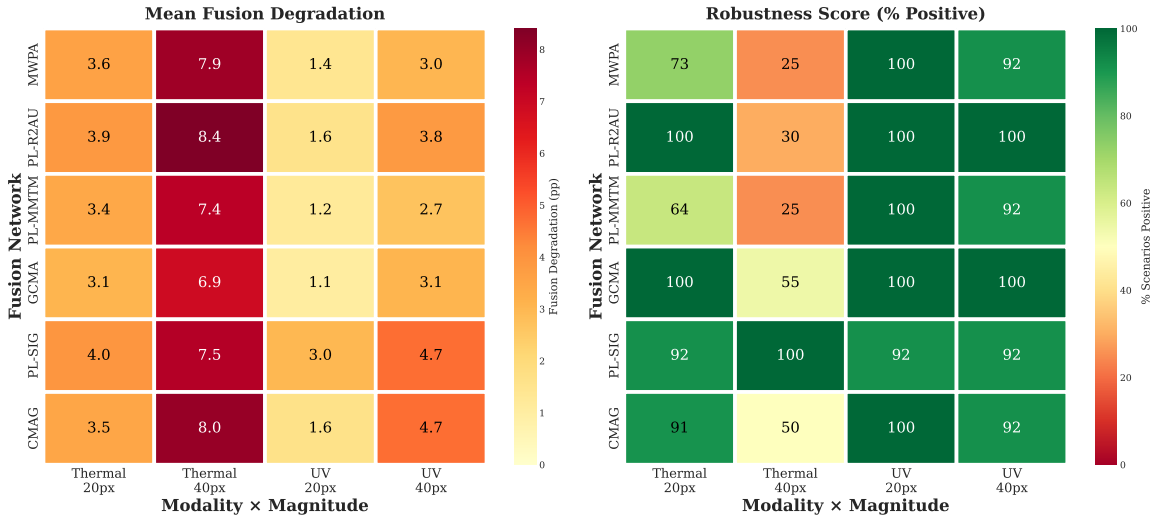


Figure 5.15: Network performance heatmap for spatial shift robustness. Each cell aggregates performance across 12 scenarios: three lighting conditions (RGB1/3/5) and four shift directions (up/down/left/right) for thermal/UV shifts at 20 px or 40 px magnitudes. **Left:** Mean fusion degradation (pp) shows performance loss—lower values (yellow) indicate robustness, higher values (red) indicate sensitivity. **Right:** Percentage of scenarios maintaining positive fusion benefit (shifted multimodal mIoU > single-modality baseline)—100% (green) indicates fusion remains beneficial across all conditions, while lower percentages (yellow/red) reveal cases where misalignment causes fusion to underperform.

shifts at 20 px maintain near-perfect fusion benefit (>95% positive) across most networks, while thermal 40 px shifts prove challenging, with positive rates dropping to 50-75% for less robust architectures.

Lighting Interaction Shift impact remains broadly stable across illumination conditions: mean degradations are 4.58 pp (RGB1), 3.87 pp (RGB3), and 3.90 pp (RGB5). The marginally higher sensitivity under underexposure (RGB1: 4.58 pp) reflects a benefit-fragility trade-off: while auxiliary modalities provide the most significant value when primary features degrade, they simultaneously become more critical to spatial correspondence, amplifying the impact of misalignment.

Summary

Complete modality removal represents the primary failure mode for decoder-level fusion, with RGB and DIN drops causing severe degradation (59.50 pp and 49.61 pp, respectively). In contrast, moderate spatial misalignment (20 px) induces minimal performance loss (2.87 pp average), validating that decoder-level fusion accommodates sensor misregistration without explicit alignment mechanisms. UV exhibits 2.4× greater spatial tolerance than thermal (3.20 pp vs. 5.20 pp mean degradation), correlating with its lower overall contribution to fusion (16.82 pp vs. 24.62 pp drop impact). Architectural robustness to perturbations (PL-MMTM, PL-R2AU superior for drops; PL-SIG superior for shifts) does not correlate with baseline accuracy (CMAG, GCMA superior), revealing an inherent trade-off between peak performance and perturbation resilience.

5.7.8 Noise Robustness Analysis

We evaluate robustness to input corruption by applying four noise types (Gaussian, salt-and-pepper, speckle, and generic additive noise) independently to each of the four modalities (RGB, DIN, T24, U8) at 14 intensity levels (0.1-5.0). This yields 16 corruption scenarios evaluated across six networks and three lighting conditions, totalling 4,032 evaluations. Results quantify both absolute performance loss (percentage points, pp) and relative degradation (%) to characterise architectural resilience to sensor noise. Class-level vulnerability patterns under noise perturbations are visualised in Appendix 5.B.3.

Overall Noise Sensitivity

Table 5.7 reports the mean loss across all modalities, noise types, intensities, and lighting conditions. PL-R2AU demonstrates the highest noise tolerance (8.80 pp mean loss, 10.6% relative degradation), followed by PL-MMTM (9.67 pp, 11.9%), with PL-SIG and GCMA occupying mid-tier positions. CMAG and MWPA exhibit the highest sensitivity (12.93 pp and 13.91 pp, respectively). Because degradation is non-linear with severity, we report anchor losses at both mild and severe levels of corruption. Between i=1.0 and i=5.0, degradation increases substantially: PL-R2AU from 6.51 to 16.74 pp (+10.23), CMAG from 8.77 to 25.59 pp (+16.82), and MWPA from 9.49 to 27.04 pp (+17.55). The widening performance spread under severe corruption confirms that architectural capacity to leverage auxiliary modalities determines robustness. Degradation is non-linear: at intensity 1.0, losses span 6.51 to 9.49 pp with modest inter-network variation (2.98 pp range); at intensity 5.0, this expands to 16.74 to 27.04 pp (10.30 pp range), demonstrating that architectural differences amplify 3.5× under severe corruption.

Table 5.7: Noise robustness summary across all modalities, noise types, and lighting. Anchor losses at mild (i=1.0) and severe (i=5.0) corruption with their difference. Lower is better.

Network	**PL-R2AU**	**PL-MMTM**	**PL-SIG**	**GCMA**	**CMAG**	**MWPA**
Mean loss (pp)	**8.80**	9.67	10.66	10.70	12.93	13.91
Relative deg. (%)	**10.6**	11.9	13.0	13.2	15.4	17.1
Mean loss @ i=1.0 (pp)	6.51	7.57	7.96	8.05	8.77	9.49
Mean loss @ i=5.0 (pp)	16.74	18.89	20.63	21.41	25.59	27.04
Δ (5.0–1.0) (pp)	10.23	11.32	12.67	13.36	16.82	17.55

Figure 5.16 presents degradation trajectories across all networks: an overall summary (left) and four modality-resolved panels (RGB, DIN, T24, U8) for direct comparison. PL-R2AU maintains the lowest or near-lowest curve across all panels, while MWPA and CMAG rise most steeply at high severities, driven primarily by their sensitivity to the UV modality U8. At mild corruption (intensity 1.0), losses span 6.51-9.49 pp (7.9-11.7% relative); at severe corruption (intensity 5.0), they reach 16.74-27.04 pp (20.2-33.1% relative).

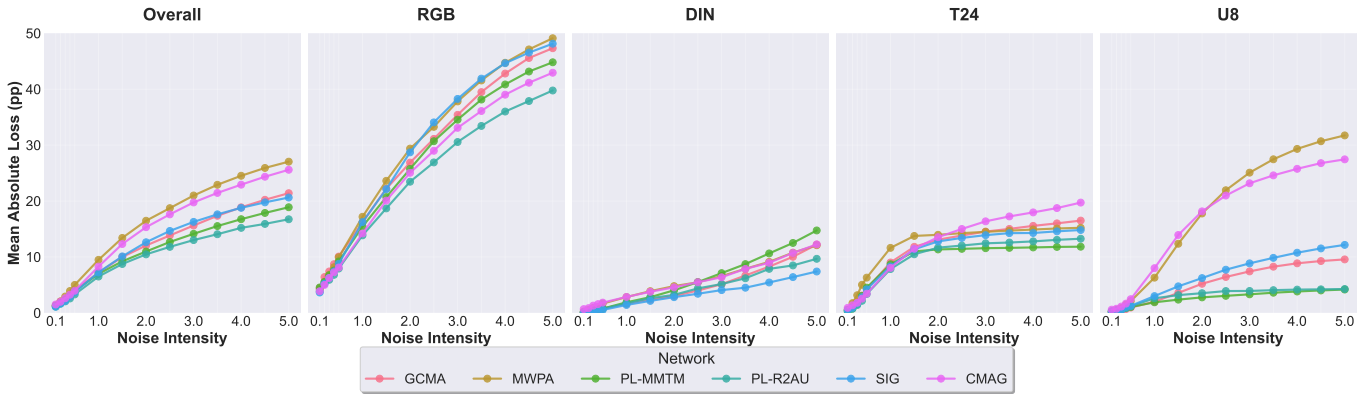


Figure 5.16: Network robustness to noise corruption across intensity levels (0.1-5.0). On the left, the overall loss averaged over all modalities, noise types, and lighting conditions. This is followed by modality-resolved panels (RGB, DIN, T24, U8) with shared y-axis (0-50 pp), enabling direct cross-modality comparison.

Modality-Specific Vulnerability

Modality vulnerability follows a clear hierarchy. Table 5.8 quantifies per-modality mean losses across networks. RGB corruption proves most damaging (20.74-25.60 pp), followed by thermal (7.94-10.59 pp), UV (2.30-14.85 pp), and depth-intensity (2.78-5.02 pp). The worst single scenario is Gaussian noise on RGB, resulting in a 28.84 pp mean loss (35.0% relative degradation) across all networks. DIN exhibits the lowest sensitivity, reflecting its auxiliary role in providing geometric cues that remain largely intact under photometric noise. UV exhibits substantial network-dependent variation: while most architectures show low UV sensitivity (2.30-5.60 pp), CMAG and MWPA demonstrate

anomalous vulnerability (13.96 and 14.85 pp, respectively).

Table 5.8: Modality-specific mean loss (pp) by network, averaged over noise types and intensities. Bold indicates best performance per modality (lowest loss).

Modality	PL-R2AU	PL-MMTM	PL-SIG	GCMA	CMAG	MWPA
DIN	3.66	5.02	**2.78**	3.96	4.94	4.84
T24	8.16	**7.94**	9.14	9.75	10.59	10.35
U8	2.63	**2.30**	5.60	4.51	13.96	14.85
RGB	**20.74**	23.41	25.14	24.58	22.23	25.60

Figure 5.17 presents scenario-level sensitivity across all noise types and intensities, confirming that RGB corruption dominates degradation, while DIN proves to be the most robust across architectures.

Figure 5.17: Scenario-level noise sensitivity heatmap (modality × noise type). Cell intensity represents mean absolute loss (pp) across networks, lighting conditions, and intensity levels. Gaussian noise on RGB constitutes the worst-case scenario (28.84 pp, 35.0% relative degradation). RGB-targeted corruption dominates across all noise types, while DIN and UV corruptions induce substantially lower degradation for most architectures.

UV-Driven Vulnerability of CMAG and MWPA CMAG and MWPA demonstrate disproportionate sensitivity to UV (U8) corruption compared to other networks, as evident in the U8 panel of Figure 5.16. CMAG suffers a 13.96 pp mean loss under U8 noise, compared to 2.30 to 5.60 pp for PL-MMTM, PL-R2AU, and PL-SIG; MWPA similarly degrades by 14.85 pp. This UV-specific vulnerability stems from CMAG's architectural design: its pre-logit attention-gating mechanism leverages UV cues more aggressively than other methods, causing unfiltered UV noise to propagate with high gain at severe intensities.

Architectural Patterns

Network rankings for noise robustness differ substantially from both baseline accuracy and spatial shift robustness, revealing distinct architectural trade-offs. PL-R2AU achieves the highest noise tolerance (8.80 pp mean degradation) despite moderate baseline performance (83.14% mIoU). Conversely, CMAG attains the highest baseline accuracy (84.18% mIoU) but exhibits greater noise sensitivity (12.93 pp mean degradation, 1.47× higher than PL-R2AU). This UV-driven vulnerability (13.96 pp loss under U8 corruption versus 2.30 to 5.60 pp for other networks) reflects CMAG's learnt fusion strategy: its hybrid attention-gating mechanism optimises auxiliary modality utilisation for clean-data performance, amplifying degradation when those channels contain noise.

The relationship between fusion strategy and perturbation tolerance reveals a fundamental trade-off. CMAG achieves the highest baseline accuracy (84.18% average mIoU) by learning to aggressively exploit auxiliary modalities; however, it consequently exhibits the highest noise sensitivity (12.93 pp mean degradation). In contrast, PL-R2AU sacrifices 1.04 pp average baseline accuracy (83.14% mIoU) yet achieves substantially superior noise resilience (8.80 pp mean degradation)—a 4.13 pp robustness improvement representing a 4.0× return on the accuracy sacrifice.
PL-MMTM demonstrates balanced performance: second-best for noise (9.67 pp mean degradation) and best for drop robustness (31.82 pp mean loss, Table 5.5), yet worst for spatial shifts (70% positive scenarios). This pattern underscores orthogonal robustness dimensions. Squeeze-and-excitation-based channel gating (PL-MMTM) enables robust handling of corrupted inputs through conservative channel weights; however, it struggles when spatial correspondence is violated, as global pooling discards the geometric structure. Conversely, spatial attention mechanisms (PL-SIG: 94% positive shift scenarios) tolerate misalignment through pixel-wise gating but can amplify auxiliary contributions, propagating noise through fusion pathways (10.66 pp mean degradation under noise).

Practical Implications

Architecture selection should account for both noise severity and modality-specific vulnerability. At mild corruption (intensity 1.0), all networks remain functional (6.51 to 9.49 pp loss, 7.9 to 11.7% relative degradation). At severe corruption (intensity 5.0), losses range from 16.74 to 27.04 pp, rendering some configurations effectively non-functional.
Prioritising RGB sensor quality yields the most significant robustness gains, as RGB-targeted noise causes mean degradation of 20.74 to 25.60 pp, versus only 2.78 to 10.59 pp for auxiliary modalities—a 2.0 to 9.2× difference. For systems that heavily leverage UV features (CMAG, MWPA), UV channel quality control becomes critical to prevent severe degradation (13.96 to 14.85 pp under UV corruption, compared to 2.30 to 5.60 pp for UV-conservative methods).
Optimal architecture choice depends on operational priorities and expected perturbation profiles. CMAG achieves the highest baseline accuracy (84.18% average mIoU) with acceptable noise tolerance at mild intensities (8.28 pp loss at intensity 1.0). However, under persistent moderate-to-severe noise (intensity > 1.5), PL-R2AU exhibits $\sim 35\%$ lower degradation than CMAG at high severity (e.g., 16.74 vs. 25.59 pp at i=5.0), while PL-MMTM is $\sim$ 26–28% lower (e.g., 18.89 vs. 25.59 pp at i=5.0). When spatial misalignment is the dominant perturbation source, PL-SIG and GCMA achieve superior shift robustness (94% and 89% positive scenarios, respectively; Section 5.7.7), though with moderately higher noise sensitivity (10.66 pp and 10.70 pp, respectively).

5.7.9 Comparative Analysis of Fusion Strategies

Baseline evaluation across six fusion architectures reveals distinct performance-robustness trade-offs under varied operational conditions. CMAG achieves the highest average accuracy with stable lighting tolerance, while adapted methods demonstrate competitive performance with reduced computational overhead. Systematic ablations reveal orthogonal robustness dimensions that inform deployment strategy:

- **Modality importance is architecturally invariant.** Drop experiments establish a consistent hierarchy (RGB > DIN > T24 > U8) across all methods, indicating that information content dominates architectural effects. However, fusion strategies exhibit substantial variation in sensitivity: PL-MMTM achieves superior drop robustness through modality isolation, while CMAG's cross-modal integration amplifies dependence on auxiliary channels (Section 5.7.7).
- **Fusion aggressiveness determines accuracy-robustness trade-offs.** CMAG's aggressive auxiliary utilisation yields the highest baseline accuracy (84.18% mIoU) but the greatest noise sensitivity (12.93 pp), while PL-R2AU's conservative fusion maintains competitive accuracy (83.14% mIoU) with superior robustness (8.80 pp). This pattern is most pronounced in UV corruption, where learnt fusion strategies determine vulnerability (Section 5.7.8).
- **Spatial robustness trades off against noise resilience.** Methods preserving spatial structure (PL-SIG, GCMA) tolerate geometric misalignment through pixel-wise correspondence, whereas global pooling architectures (PL-MMTM) discard spatial information for noise suppression. This reveals fundamental architectural constraints: spatial preservation enables alignment compensation but propagates corruption; global aggregation suppresses noise but eliminates geometric cues (Sections 5.7.7, 5.7.8).
- **Auxiliary modality utilisation determines lighting robustness.** CMAG's hybrid gating mechanism maintains stable performance under challenging illumination (RGB1: 82.54%, RGB5: 82.38%), outperforming GCMA by

3.2 percentage points under suboptimal lighting (RGB1/5: 80.49%/78.03%) through effective thermal and UV compensation when RGB features degrade. Underexposed conditions (RGB1) prove most challenging across all architectures, inducing 11% higher average degradation compared to optimal lighting (RGB3: 24.20 pp vs RGB1: 26.93 pp mean loss), confirming that challenging lighting conditions amplify dependence on auxiliary modalities.

- **Deployment context dictates optimal architecture.** For controlled environments prioritising peak accuracy, CMAG maximises performance with acceptable noise tolerance at mild intensities. Under persistent severe corruption, PL-R2AU and PL-MMTM exhibit substantially lower degradation through conservative fusion. For misalignment-dominated scenarios, PL-SIG and GCMA provide superior shift tolerance, with moderate noise sensitivity (Sections 5.7.7, 5.7.8).
- **Computational efficiency remains comparable across methods.** All architectures achieve real-time inference capability with modest overhead differences, positioning computational cost as a secondary selection criterion relative to accuracy-robustness trade-offs under expected perturbation profiles (Table 5.3).

5.7.10 Encoder- vs Decoder-Level Fusion

To contextualise decoder-level fusion within the broader multimodal segmentation landscape, we compare our pre-logit integration approach against encoder-level fusion, represented by GF-Net [4], our previous encoder-level architecture on the MM5 dataset. Both architectures employ MiT-B0 backbones and fuse RGB, depth-intensity-normals (DIN), thermal (T24), and UV (U8) modalities, enabling direct performance comparison while isolating the impact of fusion stage placement. Crucially, both networks employ Stage-Wise Intensity Fusion (SWIF) to enhance the RGB primary stream with DIN composites; the fundamental distinction lies in where auxiliary thermal and UV modalities are integrated within the feature hierarchy.

Architectural Paradigms For direct architectural comparison, we focus on sigmoid gating methods that isolate fusion stage effects: GF-Net employs encoder-level sigmoid gating, while PL-SIG implements decoder-level sigmoid gating. This comparison enables us to hold fusion mechanism complexity constant while varying only the integration stage placement, thereby isolating the architectural impact of early versus late fusion.

Encoder-level fusion (GF-Net) applies SWIF to inject DIN into RGB features at each encoder stage (Stages 1–4), establishing an enhanced RGB, DIN primary representation. Thermal and UV auxiliaries are then fused into this primary stream at every encoder stage through per-pixel sigmoid gating following CM-FRM spatial alignment [2]. This stage-wise encoder fusion propagates multimodal features through all subsequent encoder depths and the shared decoder, enabling deep cross-modal interaction at the cost of tight architectural coupling and spatial alignment dependency.

Decoder-level fusion (PL-SIG) similarly employs SWIF to enhance RGB with DIN at each encoder stage within the primary stream. However, thermal and UV modalities are processed through independent encoder-decoder pipelines, as described in Sections 5.5.3 and 5.6.

Table 5.9: Encoder- vs decoder-level fusion at VGA resolution. GF-Net refers to the SWIF-Gated (RGB+DIN+T+UV) configuration; PL-SIG is the architecturally matched decoder method (sigmoid gating only); "Decoder (best)" is CMAG from Table 5.3; "Decoder (mean)" averages the six decoder-level methods. All methods use MiT-B0 backbones and SWIF-enhanced RGB+DIN primary streams on the same MM5 test split.

Method	RGB1	RGB3	RGB5	Mean	FPS	GFLOPs
GF-Net (encoder)	84.90	88.30	84.20	85.80	55	17.3
PL-SIG (decoder)	81.27	85.31	80.29	82.29	34	84.0
Decoder (best)	82.54	87.61	82.38	84.18	31	91.7
Decoder (mean)	80.99	85.91	80.99	82.63	31–34	74.0–91.7

Quantitative Performance Comparison Table 5.9 presents accuracy and efficiency metrics for both paradigms. GF-Net achieves consistently higher baseline mIoU across all lighting conditions, with a 3.51 pp average advantage over PL-SIG (85.80% vs 82.29%). This accuracy premium is most pronounced under suboptimal illumination (RGB1:

+3.63 pp, RGB5: +3.91 pp), suggesting that stage-wise encoder integration provides stronger illumination invariance through progressive refinement across encoder depths. Under optimal lighting (RGB3), the gap narrows to 2.99 pp, indicating that both paradigms achieve comparable performance when RGB quality is high. Amongst decoder methods, CMAG achieves the highest average mIoU (84.18%), reducing the encoder advantage to 1.62 pp, though at an increased computational cost.
The efficiency disparity is substantial: GF-Net achieves 55 FPS vs PL-SIG's 34 FPS (62% higher throughput) while requiring only 17.3 GFLOPs vs 84.0 GFLOPs (79% lower computational cost). This 4.9× computational advantage stems from encoder fusion's single decoder pathway processing jointly refined features, as opposed to decoder fusion's independent per-modality encoder-decoder pipelines that operate until late integration. The parameter efficiency arises from GF-Net's shared decoder, which consumes fused encoder outputs, whereas PL-SIG maintains separate decoders for each modality stream.

Robustness Under Perturbation Ablation studies reveal contrasting vulnerability profiles between sigmoid gating paradigms. RGB removal causes severe degradation across both approaches, yet decoder fusion retains substantially higher residual performance (PL-SIG: 24.71% mean residual mIoU) compared to encoder fusion (GF-Net: 14.05% mean residual mIoU), a 1.76× advantage, as shown in the drop columns of Figure 5.18. This resilience advantage demonstrates that modality isolation in decoder fusion, where thermal and UV maintain independent processing pathways until the pre-logit stage, enables more graceful degradation when the primary sensor fails. Conversely, encoder fusion's early integration creates representational dependencies that cannot be bypassed when the base modality is unavailable. Tables 5.10 and 5.11 quantify these differences systematically.

Table 5.10: Performance degradation (pp drop and relative degradation) under complete modality removal. Values show mean mIoU loss averaged across three lighting conditions (RGB1/3/5). Decoder methods demonstrate substantially lower degradation than encoder fusion, particularly for RGB drops where PL-SIG (57.91 pp) outperforms GF-Net (71.47 pp) by 13.56 pp. The decoder mean represents the average across all six decoder-level methods.

Method	**RGB**		**DIN**		**Thermal**		**UV**		**Mean**	
	pp	**rel%**	**pp**	**rel%**	**pp**	**rel%**	**pp**	**rel%**	**pp**	**rel%**
CMAG	56.61	66.99	51.13	61.30	35.78	42.78	29.37	34.81	43.22	51.47
GCMA	61.09	74.42	51.00	62.85	35.69	42.96	20.99	25.91	42.19	51.53
MWPA	67.22	81.01	50.44	61.02	24.54	29.71	12.32	14.81	38.63	46.64
PL-MMTM	58.76	73.55	52.51	65.69	11.31	14.11	4.72	5.87	31.82	39.81
PL-R2AU	55.43	66.48	47.26	57.13	19.06	23.18	7.35	8.90	32.27	38.92
PL-SIG	57.91	69.83	45.31	55.18	21.34	26.00	26.17	31.88	37.68	45.72
Decoder Mean	59.50	72.05	49.61	60.53	24.62	29.79	16.82	20.36	37.64	45.68
GF-Net	71.47	83.44	48.04	56.54	54.08	63.32	25.66	30.07	49.81	58.34

RGB removal causes severe degradation across both approaches (Table 5.10), yet decoder fusion retains substantially more performance than encoder fusion. Converting to residual performance, PL-SIG maintains 29.91% of its baseline mIoU under RGB drop (24.71 mIoU from an 82.62 baseline), nearly 2× higher than GF-Net's 16.43% retention (14.05 from an 85.52 baseline). This resilience advantage, 13.56 pp lower degradation for PL-SIG, demonstrates that modality isolation in decoder fusion enables more graceful degradation when the primary sensor fails. The decoder mean (59.50 pp RGB degradation) outperforms encoder fusion by 11.97 pp (71.47 pp - 59.50 pp), confirming that this architectural advantage extends across all decoder variants.
DIN removal reveals comparable vulnerabilities across paradigms (decoder mean: 49.61 pp vs encoder: 48.04 pp), with PL-SIG achieving the best decoder performance (45.31 pp). This minimal 1.57 pp difference reflects the fact that both approaches rely on SWIF-enhanced RGB+DIN features established at the encoder level; removing DIN degrades this shared foundation equally, regardless of where thermal and UV auxiliaries are subsequently integrated. The comparable impact confirms that depth-intensity-normals features are equally critical to both paradigms, as geometric cues are embedded early in the feature hierarchy before the fusion stage divergence.
Thermal and UV drops expose the most striking architectural differences. For thermal removal, encoder fusion suffers severe 54.08 pp degradation, more than double the decoder mean (24.62 pp) and 2.5× worse than PL-SIG (21.34 pp). This 29.46 pp gap reveals that early thermal integration creates brittle dependencies that cascade through

Table 5.11: Performance comparison between decoder and encoder fusion under noise perturbations at intensity 2.5. Values show mIoU degradation (pp) for each modality-noise combination averaged across lighting conditions. The pivoted structure reveals modality-specific vulnerabilities: decoder fusion demonstrates superior RGB and thermal resilience, while encoder fusion's data-level DIN integration provides exceptional noise tolerance for depth features. GF-Net exhibits particularly severe thermal degradation (47.39 pp mean), exceeding decoder methods by 3.8×.

Method	Modality	Basic	Gaussian	Salt&Pepper	Speckle	Mean
CMAG (Decoder)	RGB	22.51	31.91	32.17	19.37	26.49
	DIN	3.43	6.60	8.54	5.24	5.95
	Thermal	11.09	10.45	14.87	14.12	12.63
	UV	16.10	19.04	19.51	11.90	16.64
PL-SIG (Decoder)	RGB	26.78	37.51	37.67	20.26	30.55
	DIN	2.08	3.90	4.68	3.05	3.43
	Thermal	9.85	9.48	12.82	12.43	11.14
	UV	6.52	8.42	8.03	4.28	6.81
GF-Net (Encoder)	RGB	38.57	50.33	47.93	34.35	42.80
	DIN	0.77	1.81	2.22	1.89	1.67
	Thermal	41.76	34.73	55.32	57.74	47.39
	UV	1.41	6.08	4.16	0.36	3.00

the entire encoder when disrupted. Conversely, decoder methods exhibit wide variation in auxiliary resilience: PL-MMTM (11.31 pp) and PL-R2AU (19.06 pp) demonstrate superior thermal independence through channel gating and recurrent attention, while CMAG (35.78 pp) and GCMA (35.70 pp) show higher sensitivity due to explicit cross-modal dependencies.

Noise robustness comparisons at intensity 2.5 (Table 5.11) reveal the impact of fusion stage placement across four noise types. The overall mean degradation (unweighted average across modalities) shows decoder-level PL-SIG achieving 12.98 pp, substantially outperforming encoder-level GF-Net (23.71 pp) by 10.73 pp. RGB-targeted corruption induces the most severe degradation across all methods, with encoder-level GF-Net exhibiting the highest sensitivity (42.80 pp mean) compared to decoder-level CMAG (26.49 pp) and PL-SIG (30.55 pp). Notably, CMAG demonstrates superior RGB resilience through its cross-modal attention mechanism, which adaptively redistributes representational load when primary features degrade, outperforming PL-SIG by 4.06 pp.

Thermal corruption under noise exposure exhibits catastrophic failure in encoder fusion: GF-Net suffers a mean degradation of 47.39 pp, substantially exceeding decoder-level methods (CMAG: 12.63 pp, PL-SIG: 11.14 pp) by factors of 3.8× and 4.3×, respectively. This vulnerability is particularly pronounced under salt-and-pepper (55.32 pp) and speckle (57.74 pp) noise, demonstrating that early fusion of thermal features creates brittle dependencies that catastrophically fail under severe corruption. The extreme speckle degradation (57.74 pp) represents a near-total loss of thermal information, indicating that multiplicative noise fundamentally disrupts encoder-level feature interactions. The consistent thermal advantage of decoder fusion across all noise types confirms that late integration preserves modality independence to a degree, enabling a more graceful degradation when auxiliary sensors are compromised.

Conversely, DIN exhibits exceptional noise resilience across both paradigms due to the shared SWIF mechanism, though encoder fusion achieves marginally superior tolerance (GF-Net: 1.67 pp mean degradation vs PL-SIG: 3.43 pp, CMAG: 5.95 pp). Since both architectures identically fuse DIN with RGB at each encoder stage via SWIF, producing the same enhanced RGB+DIN primary stream, this small difference arises from downstream architectural choices rather than DIN processing itself. In encoder fusion, the SWIF-enhanced stream is immediately fused with thermal and UV at each encoder stage, allowing auxiliary modalities to interact with the robust RGB+DIN representation throughout the encoder depth. In decoder fusion, the SWIF-enhanced stream propagates independently through the encoder before late fusion with auxiliaries, potentially accumulating slightly different noise characteristics. Similarly, UV corruption has minimal impact on GF-Net (3.00 pp mean), with speckle noise producing negligible degradation (0.36 pp), while decoder methods exhibit higher UV sensitivity (PL-SIG: 6.81 pp, CMAG: 16.64 pp). CMAG's attention-based fusion amplifies UV noise through explicit cross modal dependencies that propagate corruption across modalities at the decoder stage.

Across all decoder methods, mean degradation ranges from 14.3 pp to 23.0 pp, with conservative fusion strategies (PL-R2AU: 14.3 pp, PL-MMTM: 15.6 pp) achieving superior tolerance through residual connections and transfer

modules that maintain independent gradient pathways. Attention-based methods (CMAG: 21.0 pp, MWPA: 23.0 pp) trade noise resilience for baseline accuracy, as their explicit cross-modal interactions create stronger dependencies that amplify corruption effects. Both paradigms degrade substantially under severe corruption, confirming that sensor noise mitigation remains an open challenge regardless of fusion strategy. However, the 10.73 pp decoder advantage demonstrates that architectural choices significantly impact robustness margins.

Figure 5.18 visualises class-specific resilience patterns, revealing that decoder-level PL-SIG maintains more consistent per-class IoU across perturbations compared to encoder-level GF-Net. The most pronounced differences emerge under modality drops (leftmost perturbation columns), where GF-Net exhibits systematic class collapse (extensive red regions) while PL-SIG preserves moderate discrimination (yellow-green regions). This visualisation confirms that late integration's modality independence translates to more uniform degradation across semantic categories, avoiding the catastrophic class-specific failures characteristic of early fusion when primary modalities fail.

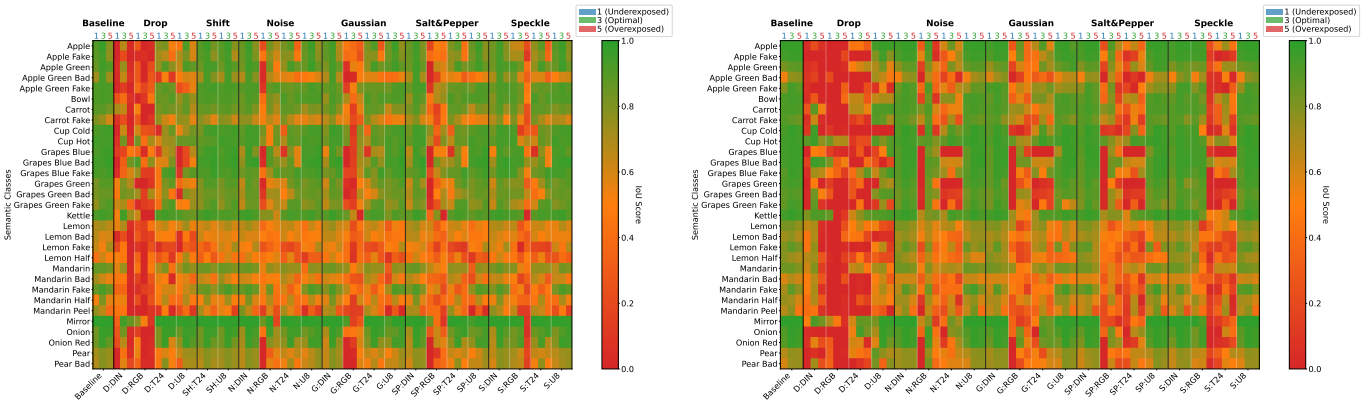
Architectural Trade-offs and Application Context The comparative analysis establishes three deployment scenarios:

(i) **Controlled environments** with geometrically aligned sensors and strict latency constraints favour encoder fusion for peak accuracy (+3.23 pp vs PL-SIG) and efficiency (+62% throughput).

(ii) **High-accuracy scenarios** tolerating modest computational overhead benefit from CMAG's hybrid fusion, achieving competitive accuracy (within 2 pp) with inherent alignment tolerance.

(iii) **Robustness-critical applications** with potential sensor failures or geometric drift require decoder-level fusion, accepting 3 pp accuracy reduction for 2.0× improved sensor failure resilience (Section 5.7.7) and intrinsic spatial tolerance (Section 5.7.7).

The choice between paradigms depends on operational constraints: encoder fusion maximises performance under assured alignment and sensor reliability, while decoder fusion prioritises robustness for challenging conditions where failures and misalignment are anticipated.



(a) PL-SIG (decoder-level): Maintains class discrimination under drops/shifts through modality isolation.

(b) GF-Net (encoder-level): Shows systematic class collapse under RGB impairment, particularly for drops.

Figure 5.18: **Class-wise robustness comparison under perturbations.** Heatmaps show per-class IoU (green=high, red=low) across baseline and perturbation conditions (Drop, Shift, four noise types) under three RGB lighting conditions (1: underexposed, 3: optimal, 5: overexposed). Decoder-level PL-SIG preserves class-specific performance more consistently than encoder-level GF-Net, particularly under modality drops (leftmost perturbation group) where early fusion creates cascading failures. The systematic difference in drop columns demonstrates decoder fusion's architectural advantage in maintaining auxiliary pathway independence.

5.7.11 Discussion

Accuracy-Robustness Trade-Off

Architectural fusion strategies exhibit distinct performance-robustness profiles. CMAG achieves the highest mean accuracy (84.18% mIoU) but demonstrates elevated noise sensitivity (12.93 pp mean degradation), while MWPA exhibits the highest noise vulnerability (13.91 pp) despite mid-range baseline performance (82.89% mIoU). However, PL-R2AU demonstrates that aggressive accuracy-robustness trade-offs are not inevitable, achieving competitive accuracy (83.14% mIoU, only 1.04 pp below CMAG) while maintaining the lowest noise sensitivity (8.80 pp)—a 4.13 pp robustness advantage over CMAG. Similarly, PL-MMTM demonstrates balanced characteristics with the lowest baseline accuracy (79.89% mIoU) yet second-best noise robustness (9.67 pp), indicating that conservative fusion strategies can maintain robustness without substantial accuracy penalties. The divergent profiles stem from learnt fusion strategies: aggressive auxiliary exploitation (CMAG, MWPA) maximises discriminative capacity under ideal conditions but amplifies vulnerability when auxiliary channels degrade, whereas selective modality utilisation (PL-R2AU, PL-MMTM) maintains robustness through conservative fusion gains.

Modality Hierarchy and Architectural Invariance

Drop ablations establish a consistent modality importance hierarchy (RGB $>$ DIN $>$ T24 $>$ U8) across all six architectures, with mean losses of 59.50 pp, 49.61 pp, 24.62 pp, and 16.82 pp, respectively, averaged across all architectures and lighting conditions. This invariance demonstrates that information content dominates architectural effects: RGB-DIN provides geometric structure and appearance, while thermal and UV contribute specialised discriminative cues for challenging classes (Section 5.7.5). However, fusion strategies substantially influence sensitivity magnitude: PL-MMTM achieves a 31.82 pp mean drop loss through modality isolation, while CMAG's cross-modal integration amplifies dependence (43.22 pp).
Thermal demonstrates greater importance than UV for fusion performance: when either auxiliary modality is removed, networks retain 57.8% mean residual mIoU without thermal compared to 65.6% without UV (averaged across all architectures and lighting conditions), a 7.8 percentage point difference that quantifies the relative contribution of each auxiliary modality to segmentation performance.

Decoder-Level Spatial Robustness

Spatial shift ablations reveal intrinsic misalignment tolerance: 20 px offsets cause only 2.87 pp mean degradation vs 37.64 pp for complete modality removal—a 13.1$\times$ difference. This resilience stems from decoder fusion's semantic-level integration, where spatial correspondence assumptions are relaxed compared to encoder-level pixel alignment. Thermal exhibits 1.63$\times$ greater shift sensitivity than UV, with vertical shifts proving most damaging.

Fusion Stage Selection and Operational Context

The encoder-decoder comparison (Section 5.7.10) reveals that the placement of the fusion stage represents a primary design decision. Encoder-level fusion maximises inference throughput (55 vs. 31–34 FPS) and computational efficiency (17.3 vs. 74.0–91.7 GFLOPs) through feature sharing, while decoder-level fusion prioritises robustness through modality isolation (retains 30.17% of baseline vs. 16.56% for encoder fusion under RGB loss). However, encoder-level fusion requires geometrically aligned inputs, imposing preprocessing overhead not reflected in the reported inference metrics. Thermal and UV images must undergo rectification and lens distortion correction. Whilst these operations execute in parallel for both modalities, conservative estimates based on typical performance for VGA-resolution thermal imagery suggest approximately 5 ms preprocessing latency per frame. Although the network maintains a 55 FPS inference capability, the mandatory preprocessing creates a 5 ms system latency and incurs additional computational costs (CPU-based rectification) before frames enter the GPU-accelerated network. In contrast, decoder-level fusion operates directly on raw, distorted sensor streams, processing frames immediately upon acquisition without the delay associated with registration preprocessing.
For controlled environments with mechanically stable sensor arrays and strict latency requirements, the accuracy advantage of encoder fusion (+1.34 pp mean across RGB1/3/5 vs. the best decoder method) may justify adoption if the 5 ms preprocessing latency and additional CPU overhead remain acceptable for the application. Conversely, for field

robotics or scenarios with mechanical vibration, sensor degradation, or calibration drift, the more graceful degradation of decoder fusion and its intrinsic tolerance to misalignment outweighs the 1 to 2 percentage point accuracy reduction. Additionally, decoder fusion eliminates the registration preprocessing requirement entirely, enabling zero-delay frame-to-prediction throughput, which is critical for reactive robotic control.

Limitations

Whilst CMAG demonstrates strong performance for unaligned multimodal fusion, several limitations warrant acknowledgement:

Computational overhead. CMAG carries a higher compute/parameter budget than the most efficient baseline (91.7 vs. 74.0 GFLOPs for PL-MMTM; ≈24% increase) and a larger parameter count (22M vs. 19M). Nevertheless, all methods sustain real-time inference (31–34 FPS). The attention path (GCMA) contributes most of this overhead.

Training complexity and memory. Using separate decoder heads per modality increases training-time memory and adds optimisation complexity. In particular, we employ per-head learning rates and multi-head supervision with loss weighting and residual warm-up. These settings improve stability and graceful degradation under missing inputs, but they also enlarge the hyperparameter search space and can make exact reproduction more sensitive to configuration.

Noise–accuracy trade-off. CMAG tends to yield larger absolute mIoU losses under severe corruptions than conservative fusion (e.g., PL-R2AU, PL-MMTM), while performing competitively at mild to moderate intensities. This reflects an explicit design choice; aggressive auxiliary utilisation maximises discriminative capacity in clean conditions but increases vulnerability when auxiliaries degrade (see Table 5.7).

Alignment tolerance bounds. Modality-level pooling in GCMA provides robustness to moderate misalignment; however, performance degrades progressively with spatial shifts. In our misregistration study, 20-pixel shifts cause modest degradation (2.61 pp average), while 40-pixel shifts induce more substantial losses (5.68 pp average; see Table 5.6). Performance under larger misalignments or non-translational distortions (e.g., rotation, scale) remains untested and warrants further investigation.

Domain scope. Results are reported on MM5 (indoor produce with controlled RGB lighting and auxiliary thermal/UV). Generalisation to other domains — e.g., outdoor scenes, autonomous driving, or medical imaging with different sensor suites — remains to be established.

Calibration and failure awareness. The present model does not include explicit confidence calibration or lightweight sensor-health cheques (e.g., dropout, drift detection). Integrating uncertainty quantification and simple failure detectors would better support safety-critical or time-critical deployments.

Future Work

Several promising directions emerge from this study. The observed accuracy-robustness trade-off motivates training strategies aimed at shifting the optimisation frontier, including corruption-aware objectives, adversarial perturbation schemes targeted at auxiliary streams (thermal/UV), and multi-objective searches that balance clean accuracy against robustness to noise, modality drop, and spatial shift. Targeted augmentation—especially spatial perturbations reflecting realistic misalignment and modality-specific noise processes—may further enhance resilience while preserving clean-data performance.

The modality-specific contributions observed suggest potential for class-adaptive fusion that selectively weights inputs by semantic context at the pre-logit stage. Extending evaluation to other multimodal domains (e.g., autonomous driving, medical imaging) would test whether the design principles identified here generalise beyond controlled inspection settings.

Computational efficiency could be improved through lightweight attention mechanisms or knowledge distillation, addressing the 24% overhead while maintaining accuracy for edge deployment. Incorporating differentiable spatial transformation networks within the decoder could extend the alignment tolerance without sacrificing the benefits of late fusion.

Finally, integrating uncertainty quantification for calibrated confidence and lightweight failure detection for sensor malfunctions (e.g., auxiliary dropout or drift) would support safer deployment in time-critical applications, particularly when operating on unaligned and optically uncorrected auxiliary streams.

5.8 Conclusion

This work presents a comprehensive investigation of decoder-level fusion strategies for multimodal semantic segmentation using unaligned RGB+DIN, thermal, and UV imagery. We propose CMAG (Cross-Modal Attention with Gated Residuals). This decoder-level fusion module combines global cross-modal attention with sigmoid-gated residuals to enable alignment-tolerant fusion without explicit geometric calibration. Through systematic evaluation of CMAG against five adapted baseline methods across three lighting conditions, we establish performance baselines, quantify modality contributions via ablation studies, and assess robustness to sensor noise across 4,032 configurations (four corruption types, fourteen intensity levels). Crucially, thermal and UV modalities are fused in their distorted form without lens correction, testing decoder-level fusion's capacity to handle realistic sensor imperfections alongside spatial misalignment.

Our findings reveal critical insights for decoder-level multimodal fusion design. Our proposed CMAG achieves the highest baseline accuracy (84.18% mIoU average, 87.61% under optimal lighting) through hybrid channel-modality attention gating at the decoder level, while maintaining moderate noise tolerance (12.93 pp mean degradation across all noise scenarios) and graceful lighting adaptation. GCMA (CMAG's attention component evaluated standalone) achieves strong optimal-lighting performance (86.72%) via cross-modal attention but exhibits increased sensitivity to suboptimal illumination (RGB1: 80.49%, RGB5: 78.03%). The baseline architectures demonstrate alternative trade-offs: PL-MMTM (adapted Multimodal Transfer Module) attains the best ablation robustness (31.82 pp mean drop loss, 1.07 pp shift loss) via squeeze-and-excitation fusion; PL-R2AU (adapted Recurrent Residual Attention U-Net) achieves the best noise resilience (8.80 pp mean loss, rising from 6.51 pp at mild corruption to 16.74 pp at severe levels) through recurrent attention mechanisms.

Ablation studies confirm a clear modality hierarchy (RGB $>$ DIN $>$ T24 $>$ U8), with RGB removal causing a 59.50 pp average degradation (72.05% relative), and DIN showing critical importance under challenging lighting conditions (up to 63.99 pp loss when removed under underexposure). Thermal and UV modalities provide specialised discriminative information essential for challenging classes (fake objects, partially decayed fruit) despite modest overall importance (24.62 pp and 16.82 pp drop impacts, respectively). Notably, decoder-level fusion of unaligned modalities demonstrates strong spatial robustness, with 20 px misalignment causing only 2.11 pp average degradation compared to 37.64 pp for complete modality removal—indicating that moderate calibration drift poses minimal risk to segmentation accuracy at the decoder level.

Noise robustness evaluation reveals that RGB-targeted corruption induces the most severe degradation (Gaussian RGB: 31.91%–37.51% impact across networks; salt-and-pepper RGB: 29.31%–38.22%), while DIN demonstrates surprising noise tolerance (3.90%–7.55% average impact) despite being the second-most-important modality. Architectural differences in noise handling amplify under severe corruption: at mild intensity (i=1.0), network losses span a modest 2.98 pp range (6.51–9.49 pp), expanding to 10.30 pp at severe levels (i=5.0: 16.74–27.04 pp)—a 3.5$\times$ amplification. Networks with higher baseline accuracy demonstrate greater noise sensitivity, while robust architectures sacrifice peak accuracy, indicating that architectural designs must balance clean-data performance against perturbation resilience.

Comparison with encoder-level fusion (GF-Net) demonstrates that this trade-off extends across fusion paradigms: encoder integration achieves higher baseline accuracy and superior efficiency (55 FPS, 17.3 GFLOPs) through early feature sharing, while decoder-level designs (31–34 FPS, 74.0–91.7 GFLOPs) demonstrate stronger resilience under sensor failure (e.g., CMAG retains 32.75% vs. 16.43% of baseline under RGB drop). However, encoder-level fusion requires per-frame geometric alignment of auxiliary modalities, typically incurring approximately 5 ms CPU preprocessing latency at VGA resolution. Decoder-level fusion eliminates this preprocessing requirement by operating directly on raw, geometrically uncorrected sensor streams, reducing system latency and simplifying deployment.

Architecture selection follows naturally from these trade-offs: encoder-level fusion (GF-Net) for controlled, high-throughput scenarios with reliable sensors where preprocessing overhead is acceptable; decoder-level fusion (CMAG, GCMA) when prioritising accuracy under modest perturbation; and robust decoder variants (PL-R2AU, PL-MMTM) when robustness is critical, with potential sensor degradation or misalignment. All decoder-level architectures achieve real-time inference (31–34 FPS), enabling practical robotic vision applications without requiring pre-aligned or geometrically corrected sensor inputs.

This work establishes comprehensive benchmarks for decoder-level multimodal fusion with unaligned, optically uncorrected inputs, providing empirical guidance for architecture selection and highlighting fundamental trade-offs in decoder-level fusion design. We acknowledge that our findings are validated exclusively on the MM5 dataset,

comprising indoor produce inspection under controlled lighting variations. Generalisation of our learnt weighting patterns to outdoor environments, medical imaging, or autonomous driving with different sensor combinations remains to be explored empirically.

List of Abbreviations

ATT	Channel and spatial dual attention
BAM	Block Attention Module
CANet	Co-Attention Network
CBAM	Convolutional Block Attention Module
CMAF	Cross-Modal Attention Fusion
CMAG	Cross-Modal Attention with Gated Residuals
CMNeXt	Cross-Modal Next
CMX	Cross-Modal X
CNN	Convolutional Neural Network
CPS	Cross-Modal Prototype Sharing modules (in TCPSNet context)
CT	Computed Tomography
D	Depth
DGFM	Dual Gate Fusion Module
DGFNet	Dual Gate Fusion Network
DIN	Depth, Intensity, and Normals (MM5 Dataset)
DSM	Digital Surface Model
ETFormer	Edge-Thermal Transformer
FCN	Fully Convolutional Network
FEM	Feature Enhancement Module
FLAIR	Fluid-Attenuated Inversion Recovery
FPS	Frames Per Second
FRM	Feature Rectification Module
FWIoU	Frequency Weighted Intersection over Union
GAP	Global Average Pooling
GCMA	Global Context Modality Attention
GF-Net	Gated Fusion Network
GMFNet	Gated Multimodal Fusion Network
GN	Group Normalisation
GT	Ground Truth
HSI	Hyperspectral Imaging
HRNet	High-Resolution Network
LF-DLM	Late Fusion Deep Learning Model
LiDAR	Light Detection and Ranging
LN	Layer Normalisation
LWIR	Long-Wave Infrared
MCAM	Multi-scale Cross Attention Module
MEFNet	Modality Expert Fusion Network
MGFNet	Multi-Gated Fusion Network
MiT	Mix Transformer
mIoU	mean Intersection over Union
MLP	Multi-Layer Perceptron
MMTM	Multimodal Transfer Module
MPA	Mean Pixel Accuracy
MRI	Magnetic Resonance Imaging
MUUFL	Multi-sensor Urban/Unstructured Fusion and Learning
MWPA	Modality-wise Parallel Attention
NIR	Near-Infrared
PA	Pixel Accuracy
PET	Positron Emission Tomography
PICNet	Prototype-based Incremental Classification Network
PL	Pre-Logit
PL-MMTM	Pre-Logit Multimodal Transfer Module
PL-R2AU	Pre-Logit Recurrent Residual Attention U-Net
PL-SIG	Pre-Logit Sigmoid Gating
PR	Primary (RGB+DIN stream)
PSPNet	Pyramid Scene Parsing Network
QSF-Net	Quality-aware Selective Fusion Network
R2AU	Recurrent Residual Attention U-Net
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
RGB-D	RGB-Depth
RGB-T	RGB-Thermal
SAR	Synthetic Aperture Radar
SE	Squeeze-and-Excitation
SGFNet	Semantic Guidance Fusion Network
SIG	Sigmoid-Gated (residuals)
SSMA	Self-Supervised Model Adaptation
SWIF	Stage-Wise Intensity Fusion
T	Thermal
T1	T1-weighted (MRI)
T1ce	T1-weighted contrast-enhanced (MRI)
T2	T2-weighted (MRI)
T24	Thermal 24-bit (MM5 Dataset)
TCPSNet	Two-stage Cross-modal Prototype Sharing Network
TH	Thermal (auxiliary stream)
U8	Ultraviolet 8-bit (MM5 Dataset)
UCTNet	Uncertainty-aware Cross-modal Transformer Network
UDFNet	Uncertainty-aware Dynamic Fusion Network
UV	Ultraviolet
VHR	Very High Resolution

Code Availability

The code used in this paper will be made publicly available at `https://github.com/martinbrennernz/MM5-Dataset` upon publication of this work.

5.A Implementation Details

5.A.1 Normalisation details

Definitions. Given pyramid features $P \in \mathbb{R}^{B\times C\times H\times W}$, LayerNorm (LN) normalises per instance over all channels and spatial positions,

$$\mathrm{LN}(P) = \gamma \odot \frac{P - \mu_{\mathrm{LN}}}{\sqrt{\sigma^2_{\mathrm{LN}} + \epsilon}} + \beta, \quad \mu_{\mathrm{LN}} = \frac{1}{CHW} \sum_{c,h,w} P_{b,c,h,w}.$$

GroupNorm (GN) partitions channels into G groups and normalises within each group:

$$\mathrm{GN}(P) = \gamma \odot \frac{P - \mu_{\mathrm{GN}}}{\sqrt{\sigma^2_{\mathrm{GN}} + \epsilon}} + \beta, \quad \mu^{(g)}_{\mathrm{GN}} = \frac{1}{(C/G)HW} \sum_{c\in\mathcal{G}_g,h,w} P_{b,c,h,w}.$$

Complexity remarks. Both LN and GN have $O(BCHW)$ work per instance; GN exposes more parallelism by reducing over groups of size C/G. For C=512, H=480, W=640, LN reduces over 157.3×10^6 elements, while GN-16 reduces over 9.83×10^6 elements per group in parallel. Empirically, after training with LN, replacing LN by GN-16 at evaluation yields small relative deviations, $\|\mathrm{LN}(P) - \mathrm{GN}_{16}(P)\|_2/\|P\|_2 \approx 10^{-2}$, while improving throughput (Section 5.6).

5.B Detailed Network Results

5.B.1 Class level results at 220 epochs

Table 5.12: Detailed class-level network comparison across lighting conditions and fusion architectures. Shows per-class IoU, overall metrics (mIoU, FIoU, MPA, PA), and mean rank scores. Best value per RGB configuration highlighted in bold. RGB1: underexposed; RGB3: optimal; RGB5: overexposed lighting. "Bad" classes are partially rotten; "Fake" classes are replicas.

Class	RGB1 CMAG	RGB3 CMAG	RGB5 CMAG	RGB1 GCMA	RGB3 GCMA	RGB5 GCMA	RGB1 PL-R2AU	RGB3 PL-R2AU	RGB5 PL-R2AU	RGB1 MWPA	RGB3 MWPA	RGB5 MWPA	RGB1 PL-SIG	RGB3 PL-SIG	RGB5 PL-SIG	RGB1 PL-MMTM	RGB3 PL-MMTM	RGB5 PL-MMTM
Apple	91.37	**96.72**	69.22	90.58	96.26	76.96	**92.81**	96.09	87.56	84.80	95.32	**90.06**	90.30	96.02	78.54	91.92	96.11	88.95
Apple Fake	89.23	93.81	78.30	89.52	93.36	82.75	**91.48**	**94.34**	88.07	85.13	92.97	**88.60**	88.78	93.64	76.54	90.22	93.46	88.55
Apple Green	82.79	**95.80**	94.01	**85.91**	94.38	84.31	80.06	94.48	93.08	82.39	95.55	**94.70**	85.68	94.52	90.72	78.34	93.55	83.31
Apple Green Bad	63.07	92.38	89.28	71.97	91.49	63.75	60.01	**94.26**	75.39	64.24	90.60	**91.52**	**72.17**	92.93	73.07	60.84	92.11	71.50
Apple Green Fake	90.58	91.90	**94.17**	**93.08**	**92.77**	91.13	91.37	92.73	85.87	90.65	90.64	91.94	90.97	92.24	91.58	92.78	90.85	89.81
Background	**99.81**	**99.86**	**99.81**	99.79	**99.86**	99.81	99.79	99.85	99.80	99.78	99.84	99.80	99.80	99.84	99.79	99.77	99.84	99.77
Bowl	**92.29**	**93.22**	91.13	90.90	92.53	**92.48**	91.87	92.13	90.03	90.01	91.97	90.42	91.93	92.15	91.80	90.23	91.48	89.35
Carrot	87.86	**90.05**	87.05	85.70	84.89	86.30	86.57	86.14	86.80	86.43	88.04	86.90	**88.23**	88.20	83.53	83.19	86.77	**87.41**
Carrot Fake	75.64	**79.74**	72.15	72.42	59.38	72.42	72.13	65.52	71.27	76.51	76.11	69.97	**80.90**	74.56	60.66	64.97	68.56	**75.70**
Cup Cold	90.00	94.98	93.58	91.74	94.66	90.94	**92.12**	94.60	**95.08**	82.90	**95.06**	94.11	91.66	93.96	94.07	91.89	92.47	93.91
Cup Hot	93.67	**94.75**	93.23	93.41	94.06	91.54	**94.46**	94.29	93.72	88.62	93.08	**93.88**	94.31	93.83	93.26	91.12	92.59	92.82
Grapes Blue	**94.24**	94.38	**95.92**	84.60	95.77	72.14	93.41	95.96	95.46	90.55	**96.10**	95.17	91.34	94.72	95.35	90.04	93.95	89.95
Grapes Blue Bad	93.71	**96.14**	96.25	93.70	95.88	96.63	93.66	93.64	**96.66**	**94.03**	93.80	96.15	93.28	94.76	96.24	93.39	93.63	94.85
Grapes Blue Fake	91.52	95.62	**95.38**	89.13	**96.24**	84.05	79.40	95.50	94.89	91.68	95.20	95.32	**92.42**	91.84	94.58	79.56	95.04	91.73
Grapes Green	81.85	89.08	**90.99**	72.32	88.19	89.27	79.80	86.52	85.76	**87.41**	88.16	90.13	81.04	**89.54**	89.84	85.87	87.86	86.60
Grapes Green Bad	**84.22**	**87.69**	84.80	81.03	87.21	84.99	81.31	83.21	85.10	82.18	86.36	**86.22**	81.55	85.57	85.45	81.94	86.24	83.81
Grapes Green Fake	82.27	**92.75**	90.54	74.74	91.19	81.49	74.11	91.83	88.63	**82.95**	92.04	85.92	75.64	87.40	**90.87**	80.51	91.36	86.41
Kettle	**92.44**	**95.87**	94.24	91.01	95.66	**95.08**	90.82	95.77	93.77	91.36	94.49	94.32	91.86	94.77	91.74	89.28	92.61	94.54
Lemon	**70.98**	**77.91**	**68.97**	66.69	75.35	66.26	66.15	72.71	66.46	66.95	72.24	64.33	68.66	71.13	65.58	69.00	69.27	63.94
Lemon Bad	**68.55**	72.04	60.28	58.89	**72.29**	54.00	60.75	67.22	57.02	60.28	64.47	53.08	66.70	58.54	**60.92**	64.80	60.94	49.74
Lemon Fake	63.60	**70.82**	**70.72**	**74.34**	60.85	58.98	59.35	63.73	61.17	56.02	62.48	58.97	63.71	52.50	70.19	68.56	41.13	54.13
Lemon Half	54.65	**67.21**	59.96	47.12	64.78	54.22	58.70	63.55	**60.70**	49.37	66.47	56.81	41.10	61.68	49.93	**60.48**	59.89	54.46
Mandarin	85.71	84.09	82.24	85.23	87.20	75.88	86.18	86.39	78.22	85.71	**87.24**	80.38	86.22	86.20	**82.32**	**87.18**	85.29	81.36
Mandarin Bad	66.98	54.36	53.85	61.04	68.31	23.14	70.22	**77.74**	38.27	69.10	64.72	43.65	72.62	62.99	**54.37**	**73.22**	58.57	45.82
Mandarin Fake	85.93	92.38	74.98	84.19	91.63	80.98	79.94	86.38	80.58	**87.29**	**92.72**	**84.73**	85.25	92.36	84.55	85.67	91.32	83.94
Mandarin Half	62.86	82.20	68.26	60.95	**83.15**	71.62	**85.48**	80.79	**73.33**	44.89	71.57	55.93	52.07	80.77	56.75	71.15	78.08	54.56
Mandarin Peel	81.90	78.74	56.35	**84.36**	**79.80**	54.39	80.16	64.76	**64.31**	53.34	60.55	33.50	75.97	66.13	38.28	63.40	66.06	31.70
Mirror	**98.75**	98.89	98.64	98.64	**98.96**	**98.68**	98.62	98.91	98.37	98.40	98.87	98.52	98.64	98.67	98.57	98.20	98.86	98.19
Onion	**94.34**	**96.63**	**95.22**	83.62	96.35	95.04	84.39	96.19	95.03	82.34	94.43	94.92	83.96	96.19	95.11	85.18	95.75	94.23
Onion Red	**93.31**	**96.18**	84.56	83.65	95.66	84.72	83.23	95.72	**95.03**	83.00	94.16	94.20	83.89	95.63	86.62	82.04	95.35	92.51
Pear	69.63	**79.01**	76.66	69.50	78.78	72.01	**73.20**	76.00	75.14	71.23	78.76	**77.07**	71.63	78.44	74.63	71.21	75.94	74.31
Pear Bad	67.43	78.19	75.56	65.77	78.17	71.06	**72.79**	75.56	73.97	68.19	**78.73**	**75.96**	68.54	78.29	73.99	68.28	75.93	74.55
Mean Rank	**10.3**	**3.1**	**9**	12.6	4.6	12.1	11.8	5.6	10.5	13.5	5.9	9.9	10.9	6.2	10.9	12.6	8.5	13.1
Overall mIoU	**82.54**	**87.61**	**82.38**	80.49	86.72	78.03	81.39	86.02	82.02	78.99	85.71	81.47	81.27	85.31	80.29	80.76	84.09	79.45
Overall FIoU	**99.37**	**99.55**	**99.37**	99.29	99.53	99.28	99.30	99.51	99.35	99.26	99.50	99.36	99.33	99.48	99.32	99.27	99.46	99.28
Overall MPA	**89.21**	**92.58**	**88.98**	88.00	91.87	85.18	88.13	91.86	88.21	86.21	91.70	87.30	88.75	91.45	87.04	87.76	90.17	86.54
Overall PA	**99.66**	**99.76**	**99.66**	99.61	99.75	99.61	99.62	99.74	99.65	99.60	99.73	99.65	99.64	99.72	99.63	99.61	99.70	99.61
FPS	31	31	31	31	31	31	34	34	34	32	32	32	34	34	34	34	34	34
Parameters	22M	22M	22M	22M	22M	22M	19M	19M	19M	19M	19M	19M	19M	19M	19M	19M	19M	19M
GFLOPs	91.7	91.7	91.7	89.2	89.2	89.2	79.0	79.0	79.0	84.0	84.0	84.0	84.0	84.0	84.0	74.0	74.0	74.0

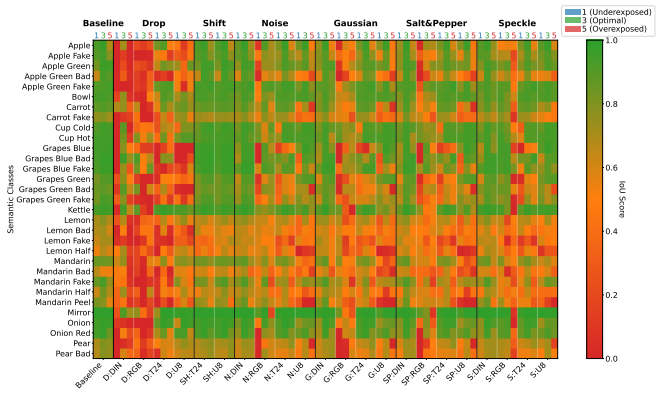
5.B.2 Network drop ablation results

Table 5.13: Drop ablation results (detailed): residual mIoU (%) after removing each modality, split by lighting condition (RGB1/3/5). Networks ordered by robustness (left to right: least to most robust). Lower residual values indicate greater modality dependence.

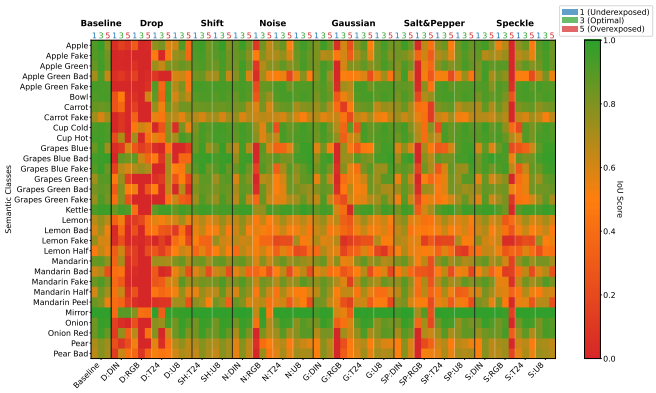
Dropped	RGB	CMAG	GCMA	MWPA	PL-SIG	PL-R2AU	PL-MMTM
RGB	RGB1	34.51	30.09	11.81	37.19	31.09	19.75
	RGB3	17.49	11.73	11.81	11.91	19.41	21.80
	RGB5	30.70	20.15	23.38	25.04	32.65	21.80
DIN	RGB1	18.54	24.01	27.63	27.89	33.40	22.05
	RGB3	57.86	45.37	42.10	56.16	52.26	32.17
	RGB5	22.73	22.86	27.63	27.89	21.98	27.91
T24	RGB1	36.30	47.97	62.48	55.85	48.07	63.23
	RGB3	61.99	31.51	65.74	72.15	77.61	74.60
	RGB5	46.89	58.68	46.83	55.85	66.57	67.92
U8	RGB1	55.48	56.48	70.57	50.03	74.44	70.61
	RGB3	57.13	54.58	70.57	60.38	77.52	78.88
	RGB5	51.93	71.24	70.57	58.98	75.72	75.17
Mean loss (pp)		**43.22**	**42.19**	**38.63**	**37.68**	**32.27**	**31.82**

5.B.3 Decoder-Level Robustness Comparison

Figure 5.19 presents comprehensive class-wise robustness heatmaps for all six decoder-level fusion architectures evaluated in this work. Each heatmap visualises per-class IoU (green=high, red=low) across baseline and perturbation conditions (Drop, Shift, four noise types) under three RGB lighting conditions (1: underexposed, 3: optimal, 5: overexposed). The layouts compare architecturally related methods: attention-based mechanisms (CMAG vs GCMA), lightweight gating versus parallel attention (PL-SIG vs MWPA), and conservative fusion strategies (PL-R2AU vs PL-MMTM). Conservative methods (bottom row) exhibit more uniform performance across perturbations, while attention-based approaches (top row) achieve higher baseline performance with increased vulnerability under severe drops. All decoder methods demonstrate superior modality isolation compared to encoder-level fusion (Figure 5.18 in the main text), with drop columns showing substantially less systematic class collapse.



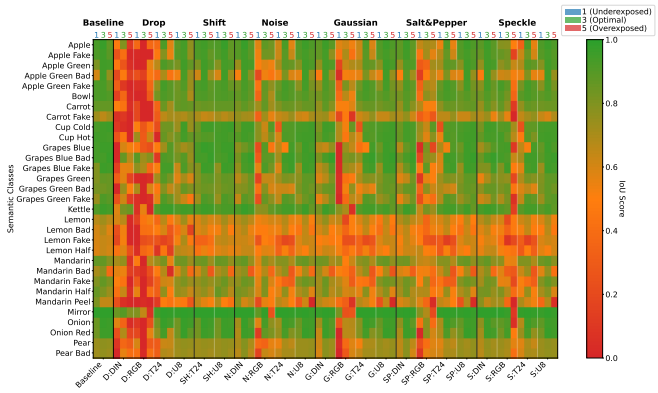
(a) CMAG



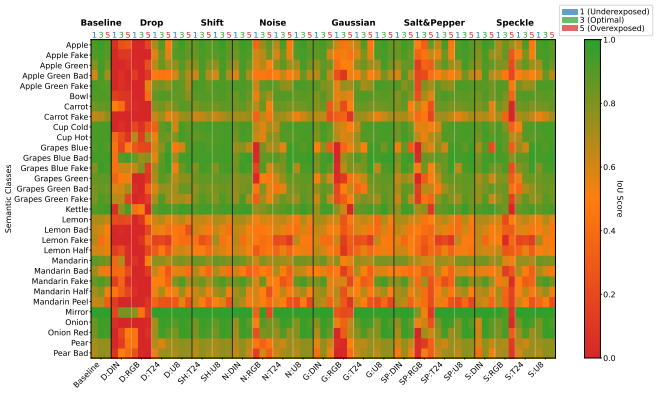
(b) GCMA

(c) PL-SIG

(d) MWPA



(e) PL-R2AU



(f) PL-MMTM

Figure 5.19: Class-wise robustness heatmaps for decoder-level fusion architectures under perturbations at intensity 2.5.

Bibliography

[1] Martin Brenner, Napoleon H. Reyes, Teo Susnjak, and Andre L.C. Barczak. Mm5: Multimodal image capture and dataset generation for rgb, depth, thermal, uv, and nir. *Information Fusion*, 126:103516, 2026.

[2] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023.

[3] Jiyuan Qiu, Chen Jiang, and Haowen Wang. ETFormer: An Efficient Transformer Based on Multimodal Hybrid Fusion and Representation Learning for RGB-D-T Salient Object Detection. *IEEE Signal Processing Letters*, 31:2928–2932, 2024.

[4] Martin Brenner, Napoleon H. Reyes, Teo Susnjak, and Andre L C Barczak. Gatedfusion-net: Per-pixel modality weighting in a five-cue transformer for rgb-d-i-t-uv fusion. *Information Fusion*, 129:103986, 2026.

[5] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020.

[6] Qiang Zuo, Songyu Chen, and Zhifang Wang. R2au-net: attention recurrent residual convolutional neural network for multimodal medical image segmentation. *Security and Communication Networks*, 2021(1):6625688, 2021.

[7] Yu Chen, Jiawei Chen, Dong Wei, Yuexiang Li, and Yefeng Zheng. Octopusnet: A deep learning segmentation network for multi-modal medical images, 2019.

[8] Zdravko Marinov, Simon Reiß, David Kersting, Jens Kleesiek, and Rainer Stiefelhagen. Mirror u-net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023.

[9] Yuhang Wang, Gongyang Li, and Zhi Liu. Semantic-guided fusion network for rgb-thermal semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7737–7748, 2023.

[10] Wenjie Lai, Fanyu Zeng, Xiao Hu, Weizhi Li, Shaobo He, Zhentian Liu, and Yadong Jiang. Mefnet: Multi-expert fusion network for rgb-thermal semantic segmentation. *Engineering Applications of Artificial Intelligence*, 125:106638, 2023.

[11] Etienne Balit and Amine Chadli. Gmfnet: Gated multimodal fusion network for visible-thermal semantic segmentation. In *Proceedings 16th the European Conference on Computer Vision*, pages 1–4, 2020.

[12] Ivica Dimitrovski, Vlatko Spasev, and Ivan Kitanovski. Deep multimodal fusion for semantic segmentation of remote sensing earth. In *ICT Innovations 2024. TechConvergence: AI, Business, and Startup Synergy: 16th International Conference, ICT Innovations 2024, Ohrid, North Macedonia, September 28–30, 2024, Proceedings*, volume 2436, page 106. Springer Nature, 2025.

[13] Kangkang Sun, Jiangyi Ding, Qixuan Li, Wei Chen, Heng Zhang, Jiawei Sun, Zhuqing Jiao, and Xinye Ni. Cmaf-net: a cross-modal attention fusion-based deep neural network for incomplete multi-modal brain tumor segmentation. *Quantitative Imaging in Medicine and Surgery*, 14(7):4579–4604, 2024.

[14] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023.

[15] Hao Zhou, Lu Qi, Hai Huang, Xu Yang, Zhaoliang Wan, and Xianglong Wen. Canet: Co-attention network for rgb-d semantic segmentation. *Pattern Recognition*, 124:108468, 2022.

[16] Xiaowen Ying and Mooi Choo Chuah. Uctnet: Uncertainty-aware cross-modal transformer network for indoor rgb-d semantic segmentation. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022.

[17] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[18] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[19] Tongxue Zhou, Su Ruan, Pierre Vera, and Stéphane Canu. A tri-attention fusion guided multi-modal segmentation network. *Pattern Recognition*, 124:108417, 2022.

[20] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020.

[21] Yongjie Guo, Feng Wang, Yuming Xiang, and Hongjian You. Dgfnet: dual gate fusion network for land cover classification in very high-resolution images. *Remote Sensing*, 13(18):3755, 2021.

[22] Liuxin Bao, Xiaofei Zhou, Xiankai Lu, Yaoqi Sun, Haibing Yin, Zhenghui Hu, Jiyong Zhang, and Chenggang Yan. Quality-aware selective fusion network for vdt salient object detection. *IEEE Transactions on Image Processing*, 33:3212–3226, 2024.

[23] Kan Wei, Jinkun Dai, Danfeng Hong, and Yuanxin Ye. Mgfnet: An mlp-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 135:104241, 2024.

[24] M. Brenner, N. H. Reyes, T. Susnjak, and A. L. C. Barczak. RGB-D and thermal sensor fusion: A systematic literature review. *IEEE Access*, 11:102667–102685, 2023.

[25] Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, 177:108635, 2024.

[26] Xiaoyan Jiang, Bohan Wang, Xinlong Wan, Shanshan Chen, Hamido Fujita, and Hanan Abd Al Juaid. Project-and-fuse: Improving rgb-d semantic segmentation via graph convolution networks. *Information Sciences*, page 122303, 2025.

[27] Tong Wang, Guanzhou Chen, Xiaodong Zhang, Chenxi Liu, Jiaqi Wang, Xiaoliang Tan, Wenlin Zhou, and Chanjuan He. Lmfnet: Lightweight multimodal fusion network for high-resolution remote sensing image segmentation. *Pattern Recognition*, 164:111579, 2025.

[28] Hui Wang, Youxiang Huang, Hao Huang, Yu Wang, Jun Li, and Guan Gui. Uncertainty-aware dynamic fusion network with criss-cross attention for multimodal remote sensing land cover classification. *Information Fusion*, 123:103249, 2025.

[29] Yongduo Zhou, Cheng Wang, Hebing Zhang, Hongtao Wang, Xiaohuan Xi, Zhou Yang, and Meng Du. Tcpsnet: Transformer and cross-pseudo-siamese learning network for classification of multi-source remote sensing images. *Remote Sensing*, 16(17):3120, 2024.

[30] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 2002.

[31] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[32] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis comprises four interlocking studies that progress from evidence synthesis to dataset construction, and finally to encoder- and decoder-level fusion architectures for robust multimodal semantic segmentation across RGB, depth, infrared intensity, thermal, and ultraviolet modalities. Three articles are published in Q1 journals, and one was submitted at the time of writing. Together, they deliver: (i) a systematic review and gap analysis of RGB-D-T fusion; (ii) a five-modality dataset and preprocessing toolkit designed for both aligned and unaligned experimentation; (iii) an efficient encoder-level architecture that establishes five-modality baselines at real-time speeds; and (iv) a decoder-level fusion family that tolerates optical distortion and cross-sensor misalignment.

Synthesis of Contributions

The systematic literature review (Chapter 2) distilled a decade of RGB-D-T fusion research, revealing that while deep learning methods surpass traditional pipelines, real-time tri-modal fusion remains challenging, and preprocessing is under-addressed. The review motivated efficient fusion beyond two modalities and highlighted practical bottlenecks—calibration, registration, and thermal/depth visualisation—that remain inadequately addressed in existing approaches.

MM5 (Chapter 3) provides the empirical foundation: a five-modality dataset with raw 16-bit depth, NIR intensity, and thermal measurements, aligned and unaligned labels via the MAR annotation remapping algorithm, and content-adaptive encodings (DTMRE for thermal; ADMRE for depth with surface normals). Baseline experiments demonstrate significant improvements from the proposed preprocessing: DTMRE thermal encoding yields approximately 6–7 pp mIoU gain over established methods, while ADMRE depth encoding with surface normals provides approximately 5–6 pp improvement over standard quantisation.

GatedFusion-Net (Chapter 4) achieves 88.3% mIoU at real-time throughput (55 fps at VGA resolution with 17.3 GFLOPs), outperforming attention-heavy baselines at a lower computational cost. The depth-intensity-normals composite raises performance under degraded RGB from 60.1% to 85.6%, confirming the complementary value of geometric cues. However, learnt gates specialise in training conditions and behave as static weights, with complete modality loss triggering up to 90.2% performance degradation.

The decoder-level study (Chapter 5) addresses this limitation through late fusion under misalignment and noise. The findings reveal a fundamental trade-off: encoder-level fusion achieves higher peak accuracy (+1.34 pp) and approximately twice the throughput, but decoder-level methods better withstand modality dropout and sensor noise. Decoder-level fusion exhibits inherent tolerance to geometric misalignment (2.61 pp degradation at 20-pixel shifts), indicating potential robustness to real-world perturbations such as sensor vibration and mounting instability.

Addressing Research Questions

This section revisits each research question posed in Chapter 1.3 and summarises how the thesis addresses them.

RQ1 *What does the current body of work on multimodal fusion reveal about datasets, calibration and registration practices, fusion strategies, and evaluation protocols, and which gaps hinder fair, reproducible benchmarking for and beyond RGB-D-T?*

The systematic review (Chapter 2) identified critical gaps in publicly available tri-modal benchmarks, modality-specific preprocessing practices, and real-time transformer-based fusion, establishing the empirical foundation for all subsequent contributions.

RQ2 *How should a reproducible capture and processing pipeline and dataset be designed to preserve sensor fidelity and support both aligned and unaligned experimentation, enabling fair comparison without enforcing a single registration choice?*

MM5 (Chapter 3) preserves raw 16-bit sensor fidelity for depth, NIR intensity, and thermal, provides both aligned and unaligned annotations via MAR, and includes systematic lighting variations, enabling a fair comparison of fusion methodologies without enforcing a single registration choice.

RQ3 *How can transformer-based encoders be adapted to integrate five modalities for dense labelling with favourable accuracy-efficiency trade-offs (mIoU vs parameters/FLOPs/latency)?*

GF-Net's (Chapter 4) stage-wise enhancement and per-pixel gating delivers real-time five-modality segmentation with 88.3% mIoU (ideal lighting with RGB3) at 17.3 GFLOPs (55 fps), demonstrating favourable accuracy-efficiency trade-offs versus channel-wise attention.

RQ4 *To what extent can a pre-logit, decoder-stage fusion scheme with per-modality heads and supervision operate directly on unaligned and optically distorted inputs, thereby eliminating explicit cross-modal geometric alignment and other preprocessing, while maintaining accuracy and interpretability? How do its components (cross-attention, sigmoid gating, stage-wise aggregation) contribute to robustness against misregistration, modality drop-outs, and environmental degradation?*

CMAG (Chapter 5) operates directly on unaligned, lens-distorted thermal and UV streams without explicit geometric calibration, achieving 87.61% mIoU (ideal lighting with RGB3). Cross-modal attention contributes baseline accuracy; sigmoid gating adds robustness; stage-wise aggregation preserves multi-scale context.

RQ5 *What is the marginal and context-dependent value of ultraviolet and thermal cues under varied illumination and noise, and how should per-pixel content-adaptive weighting compare with channel-wise mechanisms for robustness to misregistration and modality drop-out?*

Thermal cues consistently yield the most significant standalone gains, especially for detecting decay; UV aids in distinguishing synthetic replicas (Sections 4.6.4 and 5.7.5). Per-pixel sigmoid gating provides comparable accuracy to channel-wise attention at lower computational cost, though both behave as static, lighting-specific weights rather than truly adaptive mechanisms.

RQ6 *How does the choice of fusion stage—encoder-level versus decoder-level—affect accuracy, robustness to misregistration, modality dropout and sensor noise?*

Encoder-level fusion achieves +1.34 pp peak accuracy (average over three light settings) and $\sim 2\times$ throughput; decoder-level fusion provides $2.0\times$ improved sensor failure resilience and intrinsic spatial tolerance (Section 5.7.10). The choice depends on operational constraints: encoder fusion for controlled environments with reliable sensors; decoder fusion for environments where robustness to failures and misalignment is paramount.

RQ7 *How can decoder-level fusion baselines be standardised to enable a fair, reproducible comparison of alignment-free multimodal fusion methods?*

Six architecturally matched decoder-level variants (PL-MMTM, PL-R2AU, PL-SIG, GCMA, MWPA, CMAG) under a shared MiT-B0 backbone and training protocol enable controlled comparison, with code and pretrained weights released for reproducibility (Chapter 5).

Overall Contribution to Knowledge

This thesis advances multimodal perception by: (i) systematically reviewing current RGB-D-T fusion challenges and identifying actionable research gaps; (ii) creating and releasing to the public domain MM5, a five-modality dataset with raw sensor data, ground truth labels for aligned and raw data, and an accompanying preprocessing toolchain; (iii) developing an efficient encoder-level architecture that establishes the first RGB-D-I-T-UV segmentation baselines for the MM5 dataset; and (iv) developing and characterising six decoder-level fusion models as a principled approach for alignment-tolerant integration of RGB+DIN, thermal and UV imagery, thereby providing a clearer map of accuracy, robustness, and efficiency trade-offs across fusion stages.

Limitations and Future Directions

Two practical constraints remain. First, the scale of the dataset and the diversity of captures can grow further; outdoor scenes, temporal sequences, and additional object categories would strengthen generalisation claims. While DTMRE currently uses background temperature for stabilisation, production deployment would require an explicit in-frame thermal reference to maintain encoding consistency under varying ambient conditions. Second, fusion adaptability under sensor failure remains an open challenge: encoder gates behave as static weights, while decoder methods incur higher computational costs. Promising directions include uncertainty-aware fusion mechanisms that dynamically down-weight degraded streams based on estimated reliability [1, 2], and architectures that combine encoder and decoder pathways with learnt or condition-based routing to balance efficiency and robustness. Differentiable spatial alignment within late-fusion decoders may further reduce the need for explicit geometric calibration.

By connecting a targeted literature synthesis to an openly available five-modality benchmark and two complementary fusion families, this thesis advances a reproducible path towards robust multimodal perception under realistic lighting variation, geometric misalignment, and sensor degradation. It provides both empirical evidence and practical tools for future work that scales to additional modalities, broadens application scenarios, and tightens the integration between calibration, preprocessing, fusion architecture design, and reliability evaluation.

Bibliography

[1] Zheng Shao, Hai Wang, Yingfeng Cai, Long Chen, and Yicheng Li. Ua-fusion: Uncertainty-aware multimodal data fusion framework for 3d object detection of autonomous vehicles. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[2] Luke Chen, Junyao Wang, Trier Mortlock, Pramod Khargonekar, and Mohammad Abdullah Al Faruque. Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22306–22316, 2025.