Information Fusion 129 (2026) 103986

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

GatedFusion-Net: Per-pixel modality weighting in a five-cue transformer for RGB-D-I-T-UV fusion

Martin Brenner [a,*], Napoleon H. Reyes [a], Teo Susnjak [a], Andre L C Barczak [b]

[a] Massey University, Auckland, New Zealand
[b] Bond University, Gold Coast, Australia

ARTICLE INFO

Keywords:
Multimodal fusion
Thermal imaging
UV imaging
Preprocessing
Sensor fusion
Semantic segmentation
Vision transformers
Real-time fusion

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.

1. 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

* Corresponding author.
E-mail address: martin.brenner.1@uni.massey.ac.nz (M. Brenner).

https://doi.org/10.1016/j.inffus.2025.103986
Received 8 June 2025; Received in revised form 14 November 2025; Accepted 18 November 2025
Available online 26 November 2025

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 (Tables A.1 and A.2) 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.

1.1. 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.

2. 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 models [8–10], while more recent works utilise transformer-based encoders and quality-aware weighting to achieve state-of-the-art accuracy with improved efficiency [3,7,11]. 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 1 summarises representative multimodal fusion networks.

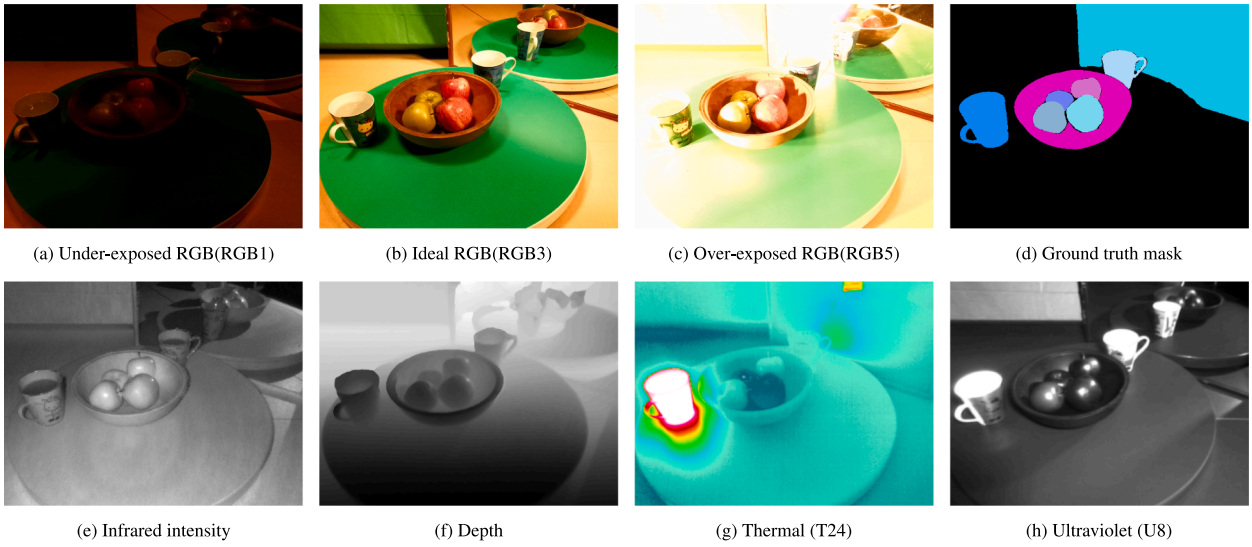
These recent advances, together with the emergence of MM5, motivate a set of unresolved research questions that underpin the present work:

- 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.

Table 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 × 480/320 × 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 × 480/320 × 240	Yes (55/91 fps)	MM5



(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)

Fig. 1. MM5 sample subset for frame 257.

3. 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 Fig. 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.

3.1. Training and evaluation data

Table 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

Table 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

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. Proposed method

Our architecture leverages a dedicated stage-wise RGB + D + I fusion module, depicted in Fig. 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 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 Fig. 2. The overall pipeline consists of: (i) a MiT-based encoder with stage-wise RGB + DIN enhancement (Section 4.1), (ii) a data-level fusion stage forming the DIN modality (Section 4.2), (iii) the SWIF module for gated cross-modal integration of Thermal and UV cues (Section 4.3), and (iv) a lightweight MLP decoder for multi-scale feature aggregation and prediction (Section 4.6).

Thermal and UV modalities are integrated using learnable sigmoid gating mechanisms, inspired by established work in both sequence modelling and multimodal fusion [14–20]. At each encoder stage k, the thermal (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 $F^{(k)}_{\text{RGB+DIN}}$, $F^{(k)}_{\text{T24}}$, and $F^{(k)}_{\text{UV}}$ 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 $G^{(k)}_{\text{T24}}$ and $G^{(k)}_{\text{UV}}$. The fused representation at each stage is thus given by:

$$F^{(k)}_{\text{fused}} = F^{(k)}_{\text{RGB+DIN}} + G^{(k)}_{\text{T24}} + G^{(k)}_{\text{UV}}. \tag{1}$$

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

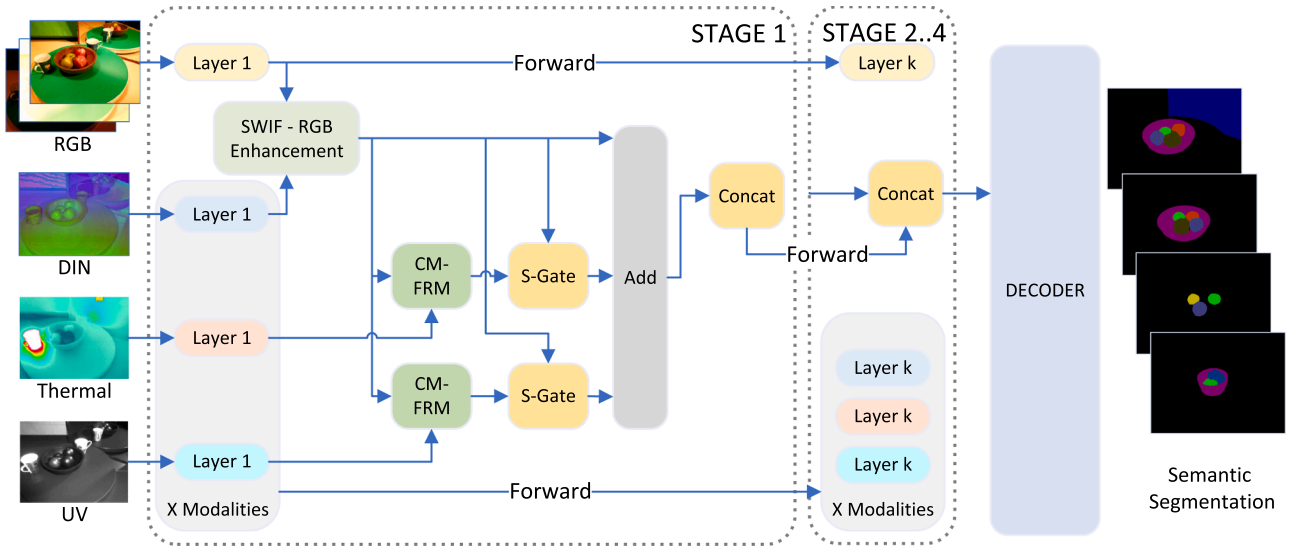


Fig. 2. Encoder-decoder pipeline with stage-wise gated multimodal fusion. For each stage, modality features are linearly projected, concatenated across modalities, and passed through a lightweight gate generator to obtain per-pixel, per-modality gates, softly normalised before fusion. The fused representation 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 Eqs. (5)–(6) for the mathematical specification.

[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.

4.1. 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 $(H_4, W_4) = (\frac{H}{32}, \frac{W}{32})$. Fig. 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:

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

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,

$$\tilde{I}^{(m,c)} = \frac{I^{(m,c)} - \mu_{m,c}}{\sigma_{m,c} + \varepsilon}, \tag{3}$$

with a small constant ε added to avoid division by zero. In practice, the dataset statistics used in Eq. (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 Fig. 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 4.7.3.

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:

$$\hat{X}_m^{(k)} = \mathrm{Linear}_{k,m}\Big(C_{m,\mathrm{in}}^{(k)} \to C_k\Big)\big(X_m^{(k)}\big), \quad m \in \mathcal{M}. \tag{4}$$

where $C_{m,\mathrm{in}}^{(k)}$ 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 $F_{\mathrm{RGB+DIN}}^{(k)}$ (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 $F_{\mathrm{RGB+DIN}}^{(k)}$
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:

$$G_m^{(k)} = \sigma\Big(\phi_m^{(k)}\Big(\mathrm{Concat}\Big(F_{\mathrm{RGB+DIN}}^{(k)}, F_m^{(k)}\Big)\Big)\Big) \odot F_m^{(k)} \tag{5}$$

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:

$$F_{\mathrm{fused}}^{(k)} = F_{\mathrm{RGB+DIN}}^{(k)} + G_{\mathrm{T}}^{(k)} + G_{\mathrm{UV}}^{(k)} \tag{6}$$

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 $\{F_{\mathrm{fused}}^{(k)}\}_{k=1}^{4}$ from all four stages forms the multi-scale output of the encoder and is forwarded to the decoder head described in Section 4.6.

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 Eq. (4), and consequently the fused outputs $F_{\mathrm{fused}}^{(k)} \in \mathbb{R}^{B\times C_k\times H_k\times W_k}$ consumed by the decoder.

Mapping into the decoder. Each fused stage output $F_{\mathrm{fused}}^{(k)}$ 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 $\mathbb{R}^{B\times N_{\mathrm{cls}}\times\frac{H}{4}\times\frac{W}{4}}$. A final bilinear upsampling by ×4 produces $\hat{Y} \in \mathbb{R}^{B\times N_{\mathrm{cls}}\times H\times W}$.

4.2. 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 Fig. 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 Appendix B.1.

4.3. 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 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 Fig. 4. At stage k, we denote the RGB and DIN feature tensors by

$$F_{\mathrm{RGB}}^{(k)} \in \mathbb{R}^{B\times C\times H\times W}, \qquad F_{\mathrm{DIN}}^{(k)} \in \mathbb{R}^{B\times C\times H\times W}, \tag{7}$$

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

$$X^{(k)} = \big[F_{\mathrm{RGB}}^{(k)}; F_{\mathrm{DIN}}^{(k)}\big] \in \mathbb{R}^{B\times 2C\times H\times W}, \tag{8}$$

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

$$\Phi^{(k)} : \mathbb{R}^{B\times 2C\times H\times W} \longrightarrow \mathbb{R}^{B\times C\times H\times W}. \tag{9}$$

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

$$C_{\mathrm{hid}} = \max\big(16, \lfloor C\cdot r\rfloor\big), \qquad r = 0.5 \text{ by default.} \tag{10}$$

The enhancement generated at this depth is

$$E^{(k)} = \Phi^{(k)}\big(X^{(k)}\big), \tag{11}$$

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

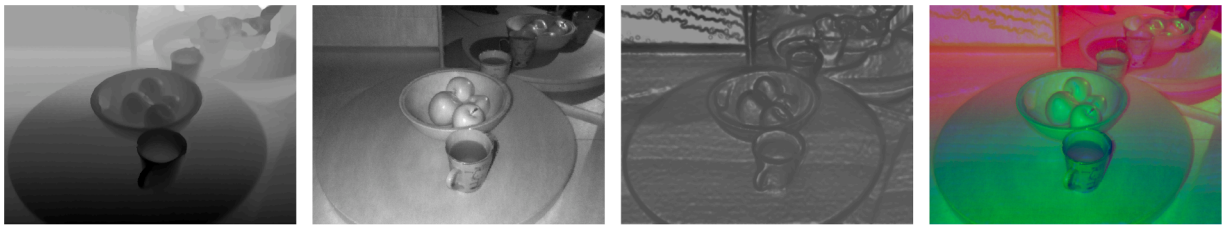
$$F_{\mathrm{RGB+DIN}}^{(k)} = F_{\mathrm{RGB}}^{(k)} + E^{(k)}. \tag{12}$$

All convolution and normalisation layers are initialised with Kaiming [22] style schemes: convolutional kernels are sampled from $\mathcal{N}\big(0, \sqrt{2/\mathrm{fan\_out}}\big)$, 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.

4.4. 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 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 Fig. 5.

While we evaluated TPS-based spatial transformers for feature alignment, we observed that the CM-FRM provided superior performance in



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

Fig. 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.

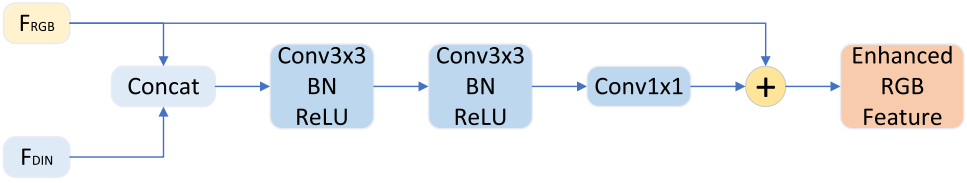


Fig. 4. SWIF (Stage-Wise Intensity Fusion) Module: The RGB and intensity feature maps are concatenated channel-wise and passed through a three-layer convolutional fusion network, producing an enhancement which is summed with the original RGB features via a residual connection to yield the stage-wise enhanced RGB representation.

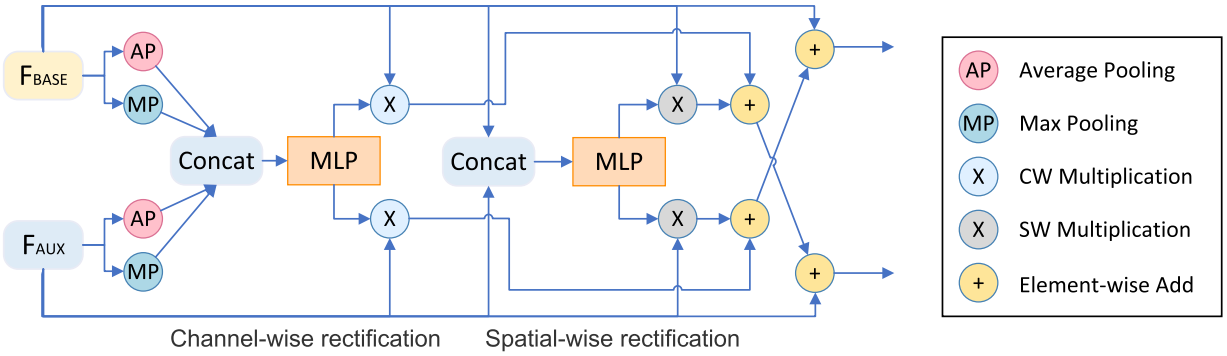


Fig. 5. The CM-FRM (Cross-Modal Feature Rectification Module [2]) architecture, detailing its channel-wise and spatial-wise feature refinement pathways for multimodal inputs.

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. 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:

$$F^{(k)}_{\text{fused}} = F^{(k)}_{\text{RGB+DIN}} + \sum_{i \in \{\text{T,UV}\}} G^{(k)}_{i}. \tag{13}$$

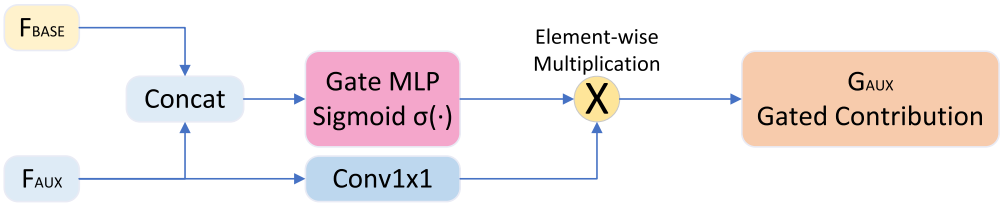


Fig. 6. Sigmoid Gate Module: The module takes a base feature map (F_{base}) and an auxiliary modality feature map (F_{aux}). For gate generation, F_{base} and F_{aux} are concatenated and processed by a Gate MLP followed by a Sigmoid function to produce a 1-channel gate. Separately, F_{aux} is processed 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}.

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 Fig. 6, which provides an overview of the key operations involved in context-aware gating and fusion.

4.6. 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 [2,25]. 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:

$$\hat{F}^{(k)} = \text{Linear}(C_k \to C)\Big(F_{\text{fused}}^{(k)}\Big), \qquad \tilde{F}^{(k)} = \text{Upsample}\Big(\tfrac{H}{4}, \tfrac{W}{4}\Big)\big(\hat{F}^{(k)}\big). \tag{14}$$

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

$$F_{\text{cat}} = \text{Concat}\big(\tilde{F}^{(1)}, \tilde{F}^{(2)}, \tilde{F}^{(3)}, \tilde{F}^{(4)}\big) \in \mathbb{R}^{B\times 4C\times \frac{H}{4}\times\frac{W}{4}}, \tag{15}$$
$$F_{\text{fuse}} = \text{Linear}(4C \to C)\big(F_{\text{cat}}\big).$$

A per-pixel classifier produces quarter-resolution logits:

$$M = \text{Linear}(C \to N_{\text{cls}})\big(F_{\text{fuse}}\big) \in \mathbb{R}^{B\times N_{\text{cls}}\times \frac{H}{4}\times\frac{W}{4}}. \tag{16}$$

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

$$\hat{Y} = \text{Upsample}\big(\times 4\big)(M) \in \mathbb{R}^{B\times N_{\text{cls}}\times H\times W}. \tag{17}$$

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 decoder consumes all four fused scales to preserve both fine detail and global semantics. Fig. 7 visualises this pipeline, showing the per-stage projection, resizing, concatenation, fusion, and final upsampling operations.

4.7. 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.

4.7.1. 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 pretrained 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

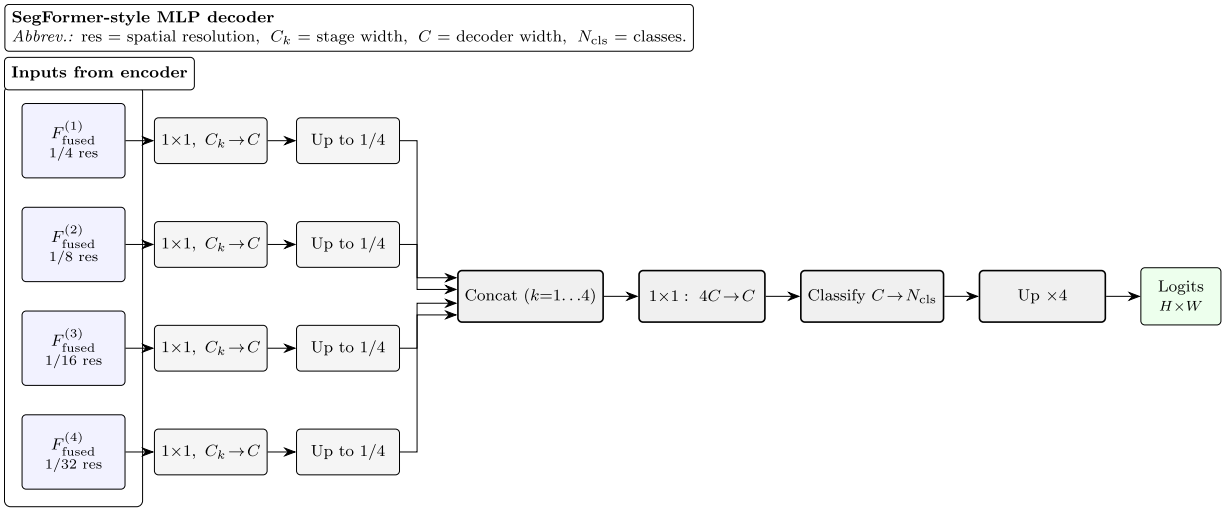


Fig. 7. MLP decoder schematic. Each fused encoder feature $F^{(k)}_{\text{fused}}$ is linearly projected to a common width C, resized to quarter resolution, concatenated across scales, fused by a pointwise linear layer and classified per pixel, then upsampled ×4 to full resolution. The design contains no attention blocks or deconvolutions, following the SegFormer and CMX practice.

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.

4.7.2. 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].

4.7.3. Training dynamics under different normalisation schemes

We compare the default ImageNet normalisation with the dataset-specific modality-wise normalisation defined in Eqs. (2) and (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. Fig. 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.

5. 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 A.1 using a MiT-B0 and Table A.2 using a MiT-B2, while an overview of the overall results is presented in Table 3.

5.1. 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,

Table 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-lIN-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

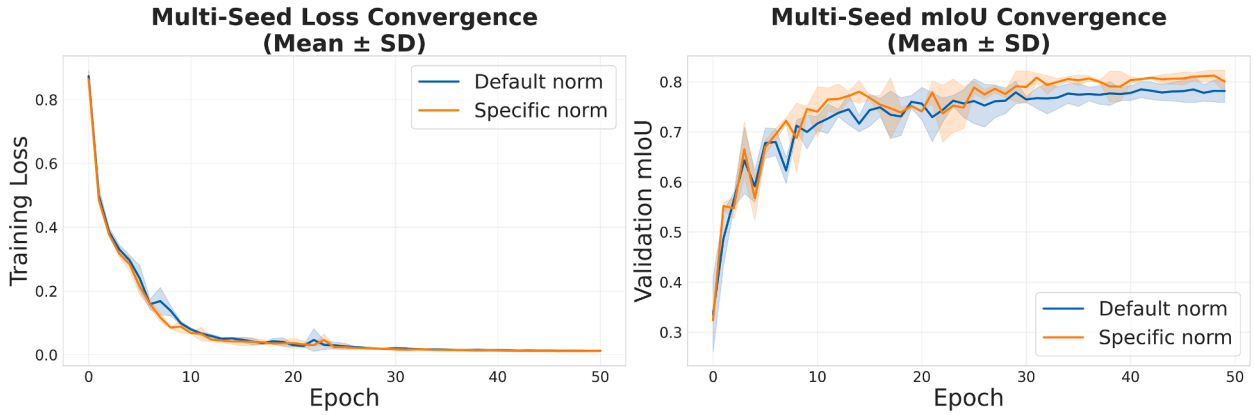


Fig. 8. Effect of normalisation on training dynamics under identical settings, with one-standard-deviation bands across three deterministic configurations. Left: mean training loss. Right: mean validation mIoU. The dataset-specific scheme shows faster convergence and improved validation accuracy.

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.

5.2. 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, 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 Fig. 9.

5.3. 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 %.
- **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.

5.4. 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

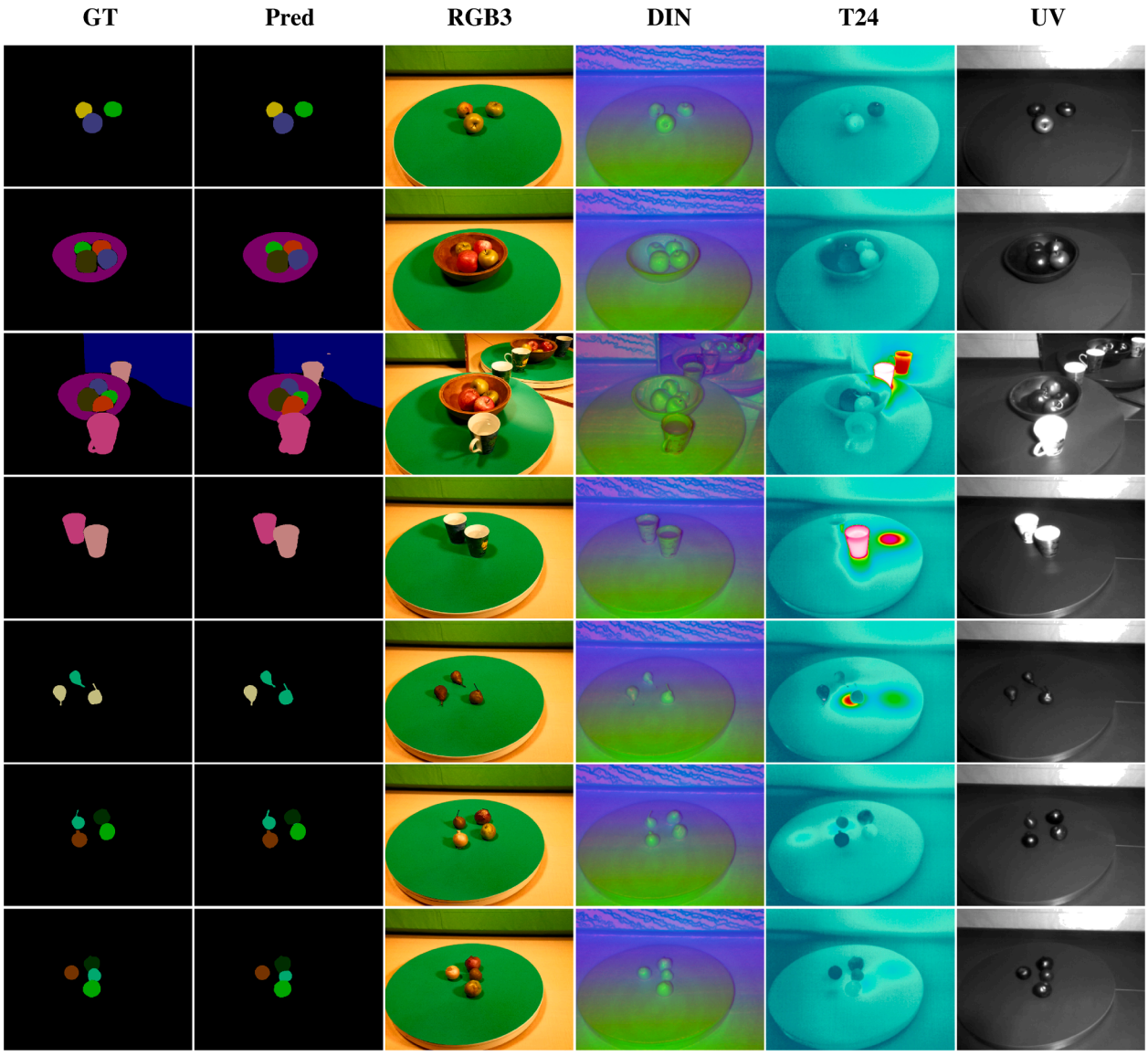


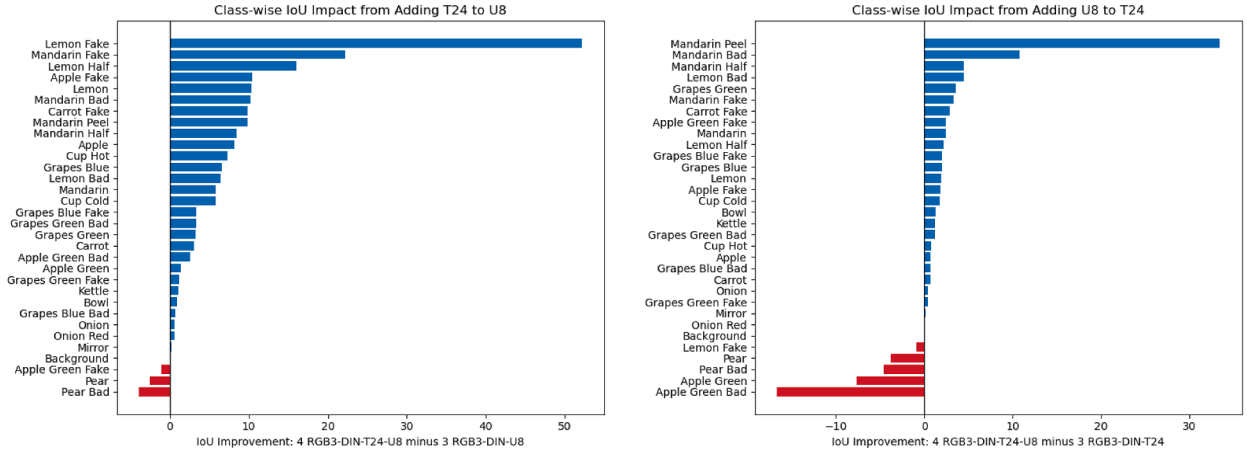
Fig. 9. Example multimodal segmentation results for seven selected frames: from left to right, ground truth mask, predicted mask, RGB, DIN (depth-intensity-normals), thermal (T24), and ultraviolet (UV). Each row corresponds to a different frame (top to bottom: 240, 250, 256, 263, 271, 289, 294).

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 Fig. 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.

5.5. 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



(a) Adding T24 to U8 fusion

(b) Adding U8 to T24 fusion

Fig. 10. Class-wise IoU impact of adding T24 (left) or U8 (right) to multimodal fusion.

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.

5.6. 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 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:

GF-Net SWIF-Gated (DIN): Depth-intensity (DIN) features are merged using the stage-wise fusion (SWIF) module as shown in Fig. 4, while all other modalities are fused via a learnt sigmoid gate as shown in Fig. 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 Fig. 11(a).

GF-Net Gated (DIN): A per-pixel sigmoid gate is applied to the fused DIN stream and all other modalities.

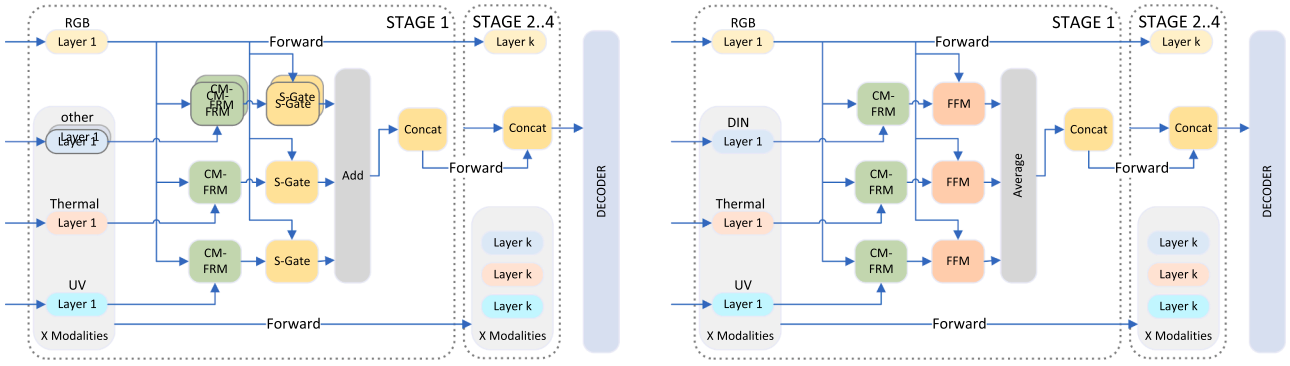
CMX FRM/FFM - PAV: We apply the cross-attention Feature Fusion Module (FFM) to each extra modality alongside RGB, then average the resulting feature maps before concatenation. In our comparisons, the 'parallel average' (PAV) strategy outperformed sequential, summation, concatenation, hierarchical, and 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 Fig. 11(b).

The quantitative results for these configurations are presented in Table 4, and a detailed class-level comparison is provided in Table A.3.

5.6.1. 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 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-



(a) GF-Net Gated (b) CMX FRM/FFM - PAV

Fig. 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).

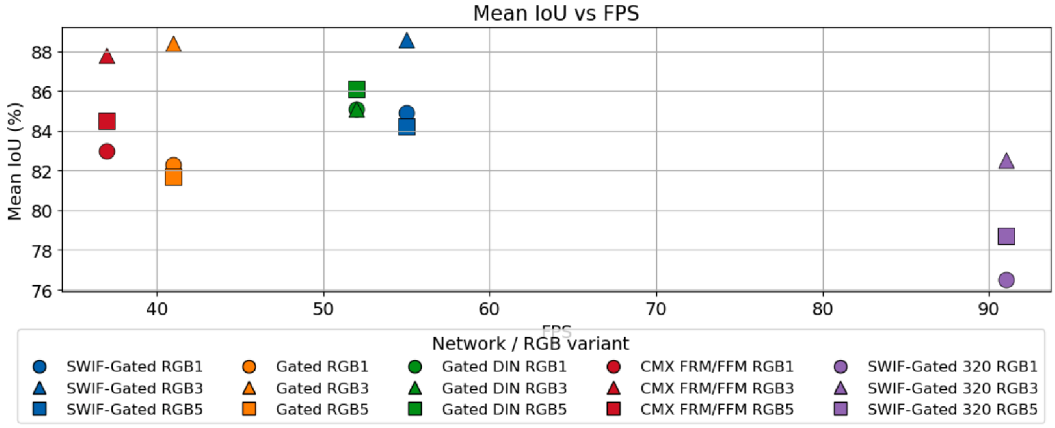


Fig. 12. Mean IoU vs FPS scatter plot of the compared networks with all modalities as shown in Table 4. RGB1 being the underexposed, RGB5 the overexposed and RGB3 the ideal lighting.

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 Fig. 12.

These findings establish a rigorous baseline for future multimodal fusion architectures on MM5 and validate the effectiveness of content-adaptive, pixel-wise gating as a robust, efficient alternative to traditional attention-based fusion. However, as detailed in Section 6, 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.

6. 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.

6.1. 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

Table 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 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	T24-U8 -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

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 capabilities.

6.1.1. *Class-specific performance analysis*

Analysis of 759 object instances across the 228 test scenes reveals persistent failure patterns for specific semantic categories. Table 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.

6.1.2. *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. Fig. 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*

Table 5
Classes with persistent high misclassification rates under full multi-modal 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 %

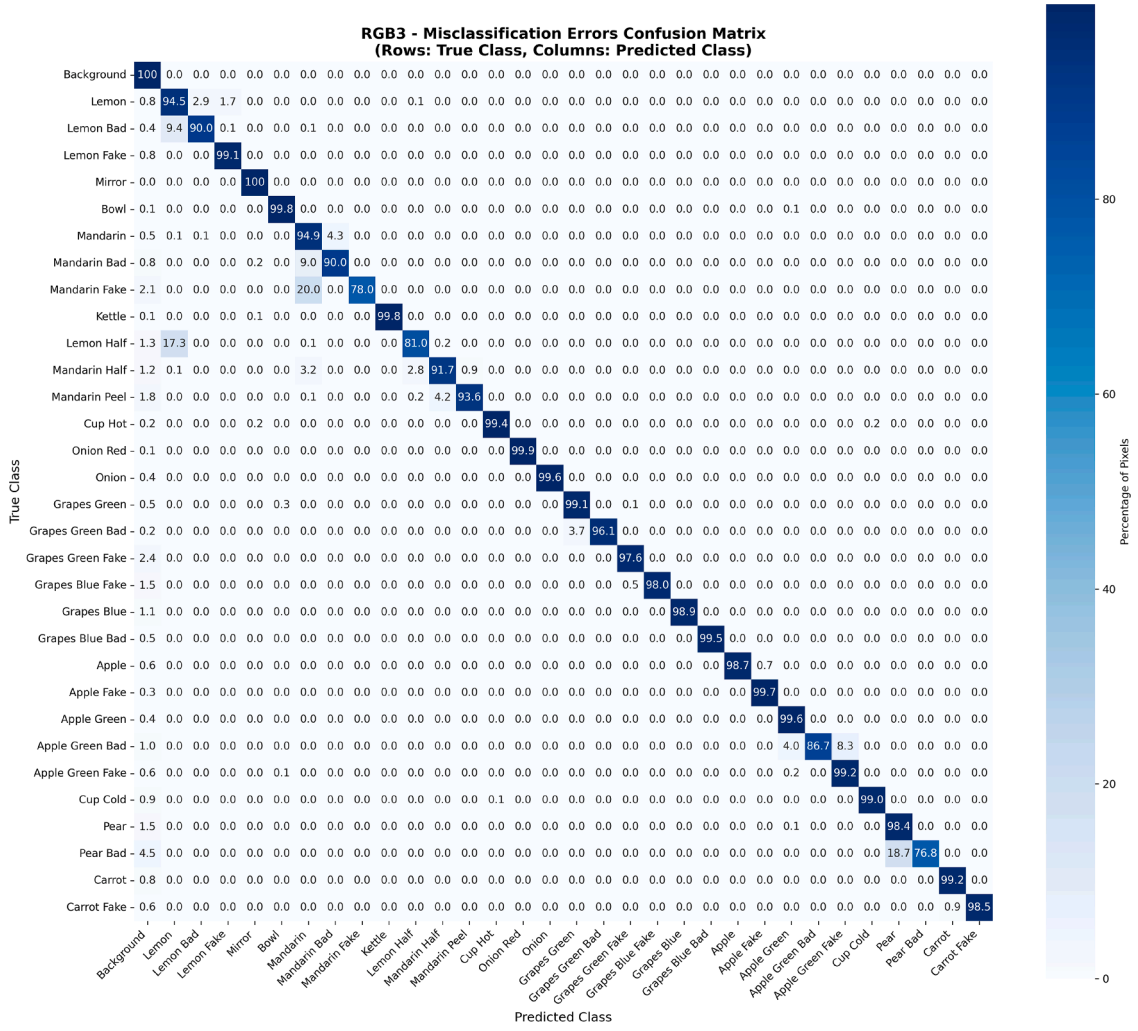


Fig. 13. Confusion matrix under optimal lighting (RGB3) showing systematic misclassification patterns between class pairs. Boundary errors within 3 pixels of edges are excluded to focus on semantic confusion rather than localisation errors.

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.
- **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

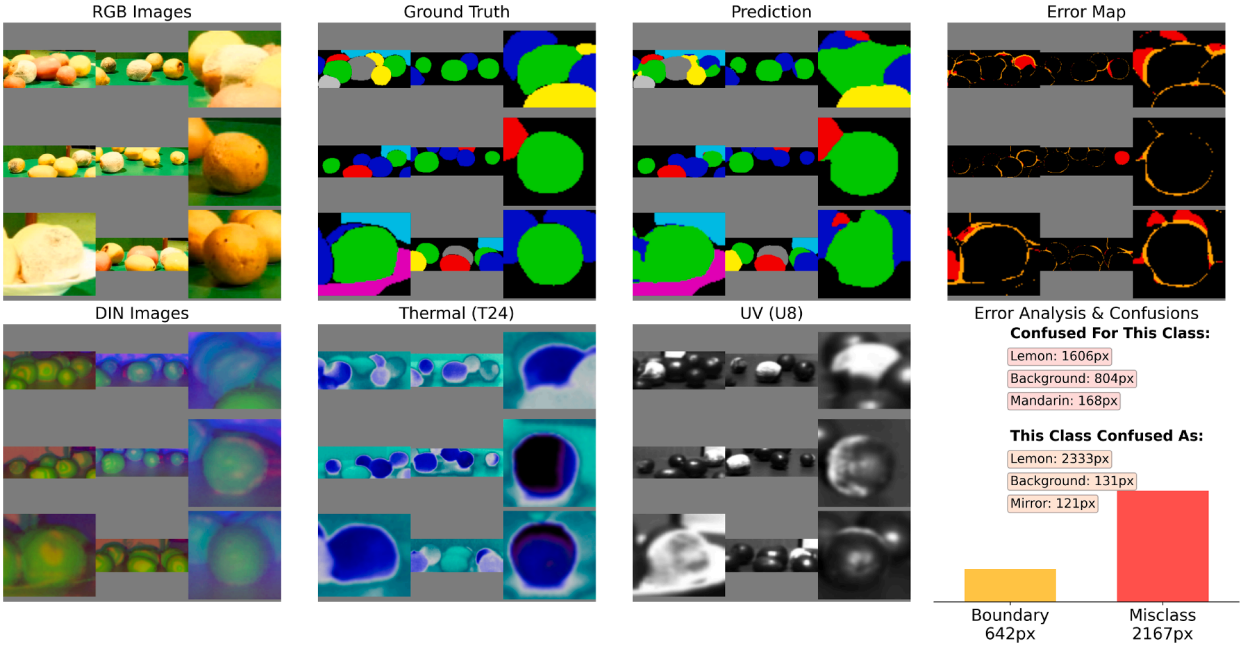


Fig. 14. Spatial error distribution for the class Lemon Bad under RGB3 conditions. The top row shows RGB input, ground truth, prediction, and error map (orange: boundary errors within 3 pixels of edges; red: misclassifications). The bottom row shows the corresponding DIN, thermal, and UV modalities. Error concentration at object boundaries and decay transitions is evident, with 642 boundary pixels versus 2167 misclassification pixels. GT: Lemon (Blue), Lemon Bad (Lime), Lemon Fake (Red), Mirror (Cyan), Bowl (Magenta), Mandarin (Yellow), Mandarin Bad (Light Grey), Mandarin Fake (Grey).

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 ±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).

6.1.3. Spatial error distribution

Pixel-level error localisation reveals that failures concentrate in predictable spatial regions rather than a random distribution. Fig. 14 illustrates the spatial patterns of boundary errors versus true misclassifications for a representative class.

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 Fig. 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.

6.2. 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 Appendix B.2. Fig. 15 illustrates the corruption types applied to the RGB3 modality.

Fig. 15. Examples of the implemented noise types. Applied to RGB3; columns show Original, Noise, Gaussian, Salt & Pepper, and Speckle.

6.2.1. 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.

6.2.2. 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, 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. Fig. 16 presents a heatmap of the class-level IoU data, and Table C.1 in Appendix C presents network-level data, including per-stage gate activations.

6.2.3. 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.

6.2.4. 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.

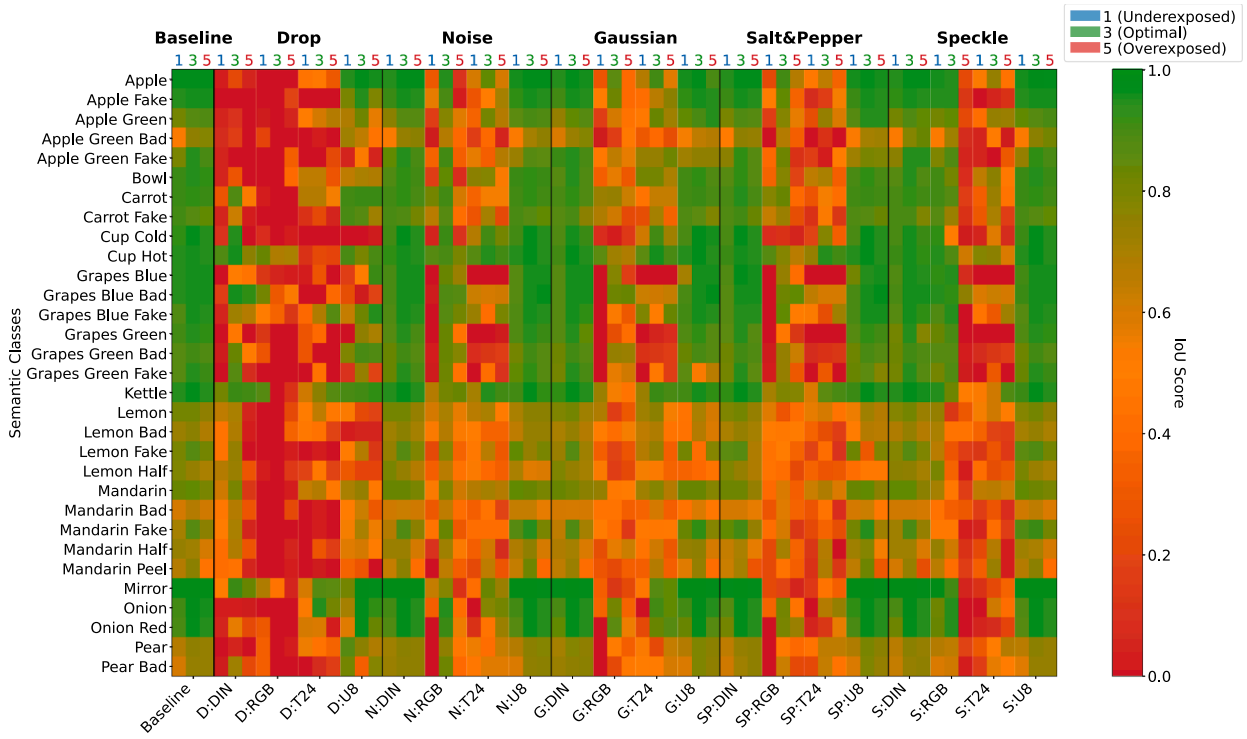


Fig. 16. Heatmap of class-wise IoU changes under drops and corruptions, red cells indicate low IoU scores and green indicate a good IoU score. Ablation Scenarios (RGB1/3/5 per scenario) D = Drop, N = Noise, G = Gaussian, SP = Salt & Pepper, S = Speckle.

6.3. 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.

7. 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- 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.

8. 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.

CRediT authorship contribution statement

Martin Brenner: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Napoleon H. Reyes:** Validation, Supervision, Resources, Project administration, Funding acquisition; **Teo Susnjak:** Validation, Supervision; **Andre L C Barczak:** Validation, Supervision.

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.

Data availability

The MM5 dataset employed in this study is publicly available, and we intend to release code examples to facilitate validation and further research upon publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Detailed network results

A.1. MiT-B0 500 Epochs

A.2. MiT-B2 250 Epochs

A.3. MiT-B0 Comparison

Appendix B. Implementation details

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.

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.

Table A.1
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

Table A.2
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

Table A.3
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

Table C.1
Comprehensive ablation analysis summary.

RGB Config	Scenario	Ablation Type	Affected Modality	mIoU 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

Appendix C. Ablation details

References

[1] M. Brenner, N.H. Reyes, T. Susnjak, A.L.C. Barczak, MM5: multimodal image capture and dataset generation for RGB, depth, thermal, UV, and NIR, Inf. Fusion 126 (2025) 103516. https://www.sciencedirect.com/science/article/pii/S1566253525005883. https://doi.org/10.1016/j.inffus.2025.103516

[2] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, R. Stiefelhagen, CMX: cross-modal fusion for RGB-X semantic segmentation with transformers, IEEE Trans. Intell. Transp. Syst. 24 (12) (2023) 14679–14694.

[3] J. Qiu, C. Jiang, H. Wang, ETFormer: an efficient transformer based on multi-modal hybrid fusion and representation learning for RGB-D-T salient object detection, IEEE Signal Process. Lett. 31 (2024) 2928–2932. https://doi.org/10.1109/LSP.2024.3465351

[4] C. Stippel, T. Heitzinger, M. Kampel, A trimodal dataset: RGB, thermal, and depth for human segmentation and temporal action detection, in: DAGM German Conference on Pattern Recognition, Springer, 2023,

pp. 18–33.
[5] N. Huang, Y. Yang, R. Xi, Q. Zhang, J. Han, J. Huang, Salient object detection from arbitrary modalities, arXiv:2405.03352 (2024). Under review.
[6] K. Song, H. Wang, Y. Zhao, L. Huang, H. Dong, Y. Yan, Lightweight multi-level feature difference fusion network for RGB-D-T salient object detection, J. King Saud Univ. Comput. Inf. Sci. 35 (10) (2023) 101702. https://doi.org/10.1016/j.jksuci.2023.101702
[7] L. Bao, X. Zhou, X. Lu, Y. Sun, H. Yin, Z. Hu, J. Zhang, C. Yan, Quality-aware selective fusion network for V-D-T salient object detection, IEEE Trans. Image Process. 33 (2024) 3212–3225. https://doi.org/10.1109/TIP.2024.3393365
[8] M. Brenner, N.H. Reyes, T. Susnjak, A.L.C. Barczak, RGB-D and thermal sensor fusion: a systematic literature review, IEEE Access 11 (2023) 102667–102685.
[9] K. Song, J. Wang, Y. Bao, L. Huang, Y. Yan, A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception, IEEE/ASME Trans. Mechatron. 28 (3) (2022) 1558–1569.
[10] H. Wen, K. Song, L. Huang, H. Wang, J. Wang, Y. Yan, Hierarchical two-stage modal fusion for triple-modality salient object detection, Measurement 218 (2023) 113180.
[11] B. Wan, X. Zhou, Y. Sun, Z. Zhu, H. Wang, C. Yan, et al., TMNet: triple-modal interaction encoder and multi-scale fusion decoder network for VDT salient object detection, Pattern Recognit. 147 (2024) 110074.
[12] A. Ozcan, O. Cetin, A novel fusion method with thermal and RGB-D sensor data for human detection, IEEE Access 10 (2022) 66831–66840. https://doi.org/10.1109/ACCESS.2022.3185402
[13] M. Brenner, N. Reyes, T. Susnjak, A. Barczak, MM5: multimodal image dataset, 2025. Dataset, https://doi.org/10.6084/m9.figshare.28722164
[14] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[15] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 933–941.
[16] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.
[17] J. Arevalo, T. Solorio, M. Montes-y Gómez, F.A. González, Gated multimodal units for information fusion, arXiv:1702.01992 (2017).
[18] Y. Cheng, R. Cai, Z. Li, X. Zhao, K. 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), 2017, pp. 3029–3037.
[19] H.R.V. Joze, A.R. Zamir, M.L. Iuzzolino, K. Koishida, MMTM: multimodal transfer module for CNN fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13286–13296.
[20] E. Balit, A. Chadli, GMFNet: gated multimodal fusion network for visible-thermal semantic segmentation, in: Proceedings 16th the European Conference on Computer Vision, 2020, pp. 1–4.
[21] Z. Guo, H. Zhang, Y. Zhang, L. Zhang, DGFNet: dual gate fusion network for land cover classification in very high-resolution images, Remote Sens. 13 (18) (2021) 3755.
[22] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level-performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034.
[23] F.L. Bookstein, Principal warps: thin-plate splines and the decomposition of deformations, IEEE Trans. Pattern Anal. Mach. Intell. 11 (6) (2002) 567–585.
[24] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, et al., Spatial transformer networks, Adv. Neural Inf. Process. Syst. 28 (2015) 2017–2025.
[25] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[26] I. Loshchilov, F. 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), https://openreview.net/forum?id=Bkg6RiCqY7.
[27] M. Tan, Q.V. Le, EfficientNet: rethinking model scaling for convolutional neural networks, in: ICML, 2019, pp. 6105–6114. arXiv:1905.11946.
[28] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: transformer for semantic segmentation, in: ICCV, 2021, pp. 7262–7272. arXiv:2105.05633.
[29] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[30] M.G. Vangel, Confidence intervals for a normal coefficient of variation, Am. Stat. 50 (1) (1996) 21–26. https://doi.org/10.1080/00031305.1996.10473533