Information Fusion 126 (2026) 103516

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

MM5: Multimodal image capture and dataset generation for RGB, depth, thermal, UV, and NIR

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

Dataset link: https://figshare.com/ndownloader/files/56868443, https://figshare.com/ndownloader/files/55555457, https://figshare.com/ndownloader/files/55555424, https://figshare.com/ndownloader/files/55555421, https://github.com/martinbrennernz/MM5-Dataset, https://doi.org/10.6084/m9.figshare.28722164

Keywords:
Multimodal dataset
Thermal imaging
UV imaging
Preprocessing
Sensor fusion
Dataset annotation

ABSTRACT

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

1. Introduction

The extraction and analysis of visual features using RGB cameras have been widely applied in computer vision across various industrial, commercial, and research domains. However, traditional RGB-based imaging is inherently limited by its confinement to the visible spectrum, making it highly dependent on external lighting conditions and susceptible to occlusions or variations in ambient illumination [1]. Expanding image capture to multiple modalities, such as depth (D), thermal (T), ultraviolet (UV), and near-infrared (NIR), can provide a more comprehensive understanding of a scene by leveraging different spectral characteristics. For example, near-infrared (NIR) imaging can penetrate certain materials or haze, and ultraviolet (UV) imaging can reveal surface details or substances not visible in RGB. Fusing such modalities with RGB yields a richer, more robust representation of the scene [2]. Studies show that multimodal combinations outperform single RGB; pairing thermal with colour imagery significantly improves pedestrian detection under difficult illumination conditions [3]. Motivated by such successes, research into multispectral perception has gained increasing attention recently due to its potential to enhance object recognition, segmentation, environmental monitoring, medical imaging, security, and robotics. Each sensor modality offers distinct advantages; for example, depth sensors provide 3D spatial information independent of texture, enabling enhanced perception of object shapes, positions, and distances; thermal cameras capture temperature differences between objects and their surroundings, ensuring reliable detection in low-light conditions and through occlusions; and UV and NIR imaging extend visibility beyond the human-perceivable spectrum, providing additional material and structural insights that can improve classification tasks. For instance, a recent comprehensive review details how the fusion of visible, NIR, and thermal imagery can dramatically improve the robustness of security systems for challenging tasks like biometric facial recognition [4], underscoring the broad scientific and

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

https://doi.org/10.1016/j.inffus.2025.103516
Received 14 April 2025; Received in revised form 29 June 2025; Accepted 7 July 2025
Available online 5 August 2025

industrial value of the sensor combination featured in our MM5 dataset. Recent advances in multimodal fusion have increasingly addressed the integration of three or more complementary sensing modalities to overcome the limitations of single-modality perception under challenging environmental conditions such as poor illumination, occlusion, and spectral camouflage. Early approaches primarily utilised convolutional neural networks (CNNs) designed for RGB, depth, and thermal infrared data, demonstrating improved robustness for salient object detection and segmentation [1,5–7]. These models often employ separate backbones for each modality, incurring higher computational costs and complexity. More recent works have adopted transformer-based architectures and attention mechanisms for more efficient and flexible fusion of multiple modalities, including RGB, depth, intensity, thermal, and ultraviolet [8–11]. Such models enable stage-wise or feature-level fusion with adaptive weighting, improving robustness against modality-specific noise and variability. Additionally, related research in remote sensing and cross-domain adaptation addresses challenges of generalising multimodal models across different geographic locations and environmental conditions, which are critical for practical deployment in diverse real-world scenarios.

However, despite the clear benefits of multimodal data fusion, research advancements in this domain are constrained by the limited availability of publicly accessible datasets that integrate multiple imaging modalities, such as RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR). As discussed in Section 2.1, many existing multimodal datasets predominantly target specific applications and frequently lack raw sensor data, resulting in restricted opportunities for exploring comprehensive multispectral fusion strategies. This limitation significantly hampers their broader applicability in multimodal fusion research. Moreover, some studies even rely on synthetic or artificially generated images [10], which fail to fully capture the complexities and challenges inherent in real-world multimodal data acquisition. Furthermore, accurate sensor calibration, precise alignment across modalities, and appropriate data preprocessing continue to pose considerable challenges, further complicating the effective utilisation of multimodal datasets in practical applications. To overcome these challenges, we introduce MM5 [12], a comprehensive multimodal dataset that systematically captures RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR) data, alongside stereo RGB-D imaging to provide complementary intensity and depth information. MM5's design specifically targets robust multimodal fusion research by offering raw 16-bit depth and thermal data, enabling researchers to apply their own denoising, enhancement, and preprocessing algorithms; accurate cross-modal alignment to ensure precise pixel correspondences across all modalities, facilitating effective multimodal data-level fusion; and annotated raw thermal and UV data generated through a novel reprojection algorithm, which remaps ground-truth labels onto original distorted thermal and UV images, significantly reducing manual labelling effort and enabling exploration of alternative fusion and alignment methods. Additionally, MM5 incorporates diverse and challenging scenarios, including reflective surfaces, hot and cold objects, and varying illumination conditions, to ensure that each sensor modality captures distinct and complementary cues. By providing both aligned and unaligned annotations and accommodating raw and preprocessed data, MM5 supports comprehensive research across multiple tiers of multimodal fusion. Researchers thus have flexibility in developing and validating methodologies within diverse processing workflows and fusion strategies. The remainder of this paper details the MM5 data acquisition framework, elaborates on sensor calibration and alignment, describes the dataset structure, and discusses potential application domains. Preliminary segmentation experiments using the Segformer-based CMX model further demonstrate the dataset's utility, highlighting the impact of different modality combinations, preprocessing approaches, and lighting variations on robust multimodal fusion.

1.1. Key contributions

The primary contributions presented in this paper are:

1. **A comprehensive multimodal dataset 8:** MM5 [12] integrates RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR) imagery, addressing existing gaps in publicly available multimodal datasets and enabling extensive multimodal fusion research.
2. **A robust and reproducible data acquisition and processing pipeline:** The pipeline systematically addresses consistent ambient conditions 4.1, sensor calibration 6 and modality alignment Appendix A.3 in a labelling and post-processing pipeline 7, significantly enhancing dataset usability and facilitating accurate multimodal analysis.
3. **MAR: Multimodal Annotation Remapping 7.1:** A novel algorithm that reprojects ground-truth annotations onto original distorted thermal and UV images, enabling flexible experimentation with both aligned and unaligned data without the overhead of fully manual annotation.
4. **DTMRE: Deterministic Thermal Multi-Resolution Encoding 9.1:** An algorithm to provide stable 24-bit colour representation and enhanced thermal resolution, particularly in regions of interest. DTMRE transforms raw thermal data into a visually informative format, facilitating better feature extraction in thermal modality-based tasks.
5. **ADMRE: Adaptive Depth Multi-Resolution Encoding 9.2:** A depth preprocessing technique that adaptively enhances depth resolution in areas exhibiting significant spatial changes or regions of interest, improving the utility of depth information in multimodal fusion.

2. Related work

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

fusion network that integrates colour images with depth maps using multi-neighbourhood graph convolutions, demonstrating significant improvements in indoor scene classification. These recent studies consistently report that multimodal input yields superior performance over single modalities, particularly under challenging conditions such as low lighting, camouflage, or sensor noise.

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

2.1. Existing multimodal datasets

Until recently, most datasets were limited to two modalities, such as NYU Depth (RGB, Depth) [23] or KAIST Multispectral (RGB, Thermal) [24]. Here, we highlight efforts that combine three or more sensor streams. Palmero et al. (2016) introduced one of the first triple-modal datasets, the VAP Trimodal People Segmentation Dataset [13], featuring approximately 11.5k frames (5.7k labelled) of indoor scenes with synchronised RGB, Kinect depth, and thermal infrared footage. Although spatially calibrated and annotated with per-pixel human masks, the dataset is limited to three static scenes with few participants, underscoring the need for broader-scale trimodal datasets. More recently, Stippel et al. (2023) expanded significantly on this concept with the TRISTAR dataset [14]. TRISTAR includes 15,618 frames of tri-modal RGB, depth, and thermal streams across ten distinct indoor environments, along with semantic segmentation and temporal action detection annotations. For face anti-spoofing, Zhang et al. (2019) created the large-scale dataset CASIA-SURF capturing RGB, depth, and near-IR video streams of 21k sequences from 1000 individuals [25]. This dataset provides annotations distinguishing real and spoofed faces, facilitating multimodal anti-spoofing research. In gesture recognition, the MGR-Dark dataset by Shi et al. (2024) comprises over 31k video clips of dynamic hand gestures recorded simultaneously in RGB, depth, and IR modalities under varying lighting conditions [16]. A dataset targeting autonomous driving scenarios is InfraParis, introduced by Franchi et al. (2024), it contains 7301 street-view images with RGB, thermal IR, and depth modalities [26], alongside detailed semantic segmentation and bounding box annotations. Similarly, Baltaxe et al. (2023) presented a polarimetric dataset featuring synchronised polarimetric, RGB, and LiDAR imagery captured across diverse road conditions [27]. In robotics, comprehensive multi-sensor datasets such as the Multi-modal and Multi-scenario SLAM Dataset for Ground Robots (M2DGR) [28], incorporating RGB, thermal, and event-camera imagery, LiDAR scans, IMU measurements, and GPS information, have significantly advanced research in robot localisation and SLAM applications. This trend extends into specialised agricultural robotics, exemplified by datasets such as CitrusFarm by Teng et al. (2023) [29], which combines RGB stereo, depth, monochrome, NIR, and thermal imagery in seven extensive sequences (1.3 TB total) captured using a custom multi-sensor rig, specifically designed for crop monitoring and robotic navigation tasks. Similarly, FieldSAFE (Kragh et al., 2019) [30] provides approximately two hours of ROS-bagged RGB stereo (including 360° panoramic), thermal, LiDAR, and radar data, annotated for obstacle detection in farming environments. The Fruity dataset (Abdulsalam et al., 2023) [31] further addresses precision agriculture with 11,065 annotated RGB, depth, and thermal images alongside pose data. Depth images are provided in 16-bit format, whereas thermal data are available in 8-bit. Conversely, Navarro et al. (2022) released a novel ground-truth multispectral image dataset of grape berries (grape Berries), including weight, anthocyanin, and Brix index measures, designed for machine learning applications [32]. It offers high-resolution 37-band VIS–NIR multispectral images for food quality assessment but lacks explicit annotations. For salient object detection, the VDT-2048 dataset (Song et al., 2022) [33] presents 2048 RGB, depth, and thermal image triplets. While it includes annotated ground truths, the thermal imagery is limited to 8-bit AGC-processed data and. We have summarised these key publicly available multimodal datasets, containing more than two imaging modalities, in Table 1. MM5 distinguishes itself from existing comparable datasets such as VDT-2048 not only through its inclusion of five spectral modalities and raw 16-bit data, but also by offering both aligned and unaligned annotations and a systematic set of eight controlled lighting conditions for RGB and three controlled settings for UV, consistently applied across all scenes. In contrast, VDT-2048 applies lighting variation sporadically and only to a subset of scenes, thus limiting its utility for comprehensive illumination-invariant fusion research.

While these existing multimodal datasets have demonstrated the advantages of integrating multiple imaging modalities, each exhibits certain constraints. Typically, these limitations include a restricted number of modalities, the absence of simultaneous provision of both aligned and unaligned annotated data, or a lack of raw sensor data, restricting research flexibility. The MM5 dataset introduced in Section 8 addresses these shortcomings by offering a comprehensive multimodal collection comprising RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR) imagery. MM5 uniquely supplies raw and preprocessed depth and thermal data alongside aligned and unaligned annotations, enabling extensive exploration across all multimodal fusion and preprocessing levels. Furthermore, the dataset includes specifically designed scenes containing real fruits, plastic fruit replicas, partially rotten produce, and other challenging elements like reflective surfaces, providing distinct sensor cues for effective feature extraction and fusion. Additionally, hot and cold objects introduce temperature contrasts, reflective objects challenge depth sensing, and varying illumination conditions further enrich the dataset, empowering researchers to investigate various fusion strategies and assess modality-specific challenges. Thus, MM5 is a robust foundation for advancing multimodal fusion research and methodology development, as detailed in Section 8.

2.2. Multi-resolution thermal encoding

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

In summary, prior work demonstrates a clear division: static linear approaches reliably maintain temperature references but often suffer from poor contrast, while adaptive methods, whether histogram-based or multi-scale, dramatically improve detail at the cost of interframe consistency. Recent multi-resolution methods [40,46–48] aim for a

Table 1
Comparative summary of publicly available multimodal datasets with more than two imaging modalities or distinct wavelengths.

Dataset (Year)	Sensors/Specs	D Res	T Res	# Samp.	Domain	Modalities present									Anno		Data format						
						V	D	T	U	N	L	R	3	I	A	U	V	D	T	U	N	L	R
VAP Trimodal People Segmentation Dataset (2016) [13]	Kinect v2, FLIR A3	512 × 424	320 × 240	11.5k frames	Indoor segmentation	✓	✓	✓	–	–	–	–	–	–	✓	–	8	16	8	–	–	–	–
TRISTAR (2023) [14,34]	Intel D435, FLIR Lepton	640 × 480	160 × 120	15.6k frames	Action segmentation	✓	✓	✓	–	–	–	–	–	–	✓	–	8	16	8	–	–	–	–
CASIA-SURF (2019) [25]	Intel SR300	640 × 480	–	21k clips	Face anti-spoofing	✓	✓	–	–	✓	–	–	–	–	✓	–	8	8	–	–	8	–	–
MGR-Dark (2024) [16]	Kinect v2 + IR	512 × 424	–	31k clips	Gesture recognition	✓	✓	–	–	✓	–	–	–	–	✓	–	8	16	–	–	8	–	–
InfraParis (2024) [26]	Stereo V, FLIR Tau2	–	640 × 512	7.3k images	Autonomous driving	✓	✓	✓	–	–	–	–	–	–	✓	–	8 S	disp	8	–	–	–	–
Polarimetic Imaging for Perception (2023) [27]	Polaris, DSLR, Velodyne	–	–	12.6k frames	Driving perception	✓	–	–	–	–	✓	–	–	–	✓	–	8	–	–	–	–	pts	–
M2DGR (2022) [28]	Multi-cam, FLIR Boson	–	640 × 512	36 sequences	Robotics SLAM	✓	–	✓	–	–	✓	–	–	✓	✓	–	8	–	8	–	–	pts	–
CitrusFarm (2023) [29]	Stereo V-D, FLIR T, NIR	–	640 × 512	7 sequences	Agriculture (robotics)	✓	✓	✓	–	✓	–	–	–	✓	✓	–	8 S	disp	8	–	8	–	–
FieldSAFE (2019) [30]	Multisense V, FLIR A65, Velodyne, Radar	–	640 × 512	2 h ROS data	Agriculture (obstacles)	✓	✓	✓	–	–	✓	✓	✓	✓	✓	–	8 S	disp	8	–	–	pts	rd
Fruity (2023) [31]	Stereo V-D, Thermal	640 × 480	640 × 480	11.0k images	Fruit picking	✓	✓	✓	–	–	–	–	–	–	✓	–	8	16	8	–	–	–	–
Grape Berries (2022) [32]	Lab (LED)	–	–	1.3k images	Food Quality	–	–	–	–	✓	–	–	–	–	–	–	–	–	–	–	16	–	–
KAIST (2018) [35]	Flea3, FLIR A655Sc, Velodyne	–	640 × 480	95k images	Driving	✓	–	✓	–	–	✓	–	–	–	✓	–	8 S	s	8	–	–	pts	–
WMCA (2020) [36,37]	Basler acA1920/acA1921, Xenics Gobi-640, Intel D415	640 × 480	640 × 480	2904 sequences	PAD	✓	✓	✓	–	✓	–	–	–	–	✓	–	8	8	8	–	–	–	–
VDT-2048 (2022) [33]	Kinect v2, T640	512 × 424	640 × 480	2048 images	SOD	✓	✓	✓	–	–	–	–	–	–	✓	–	8	16	8	–	–	–	–
MM5 (2025) [12]	Azure Kinect, iRay Micro 640, Sony XC-EU50	1024 × 1024	640 × 512	324 scenes; 2592 images	SOD; Multimodal Fusion	✓	✓	✓	✓	✓	–	–	–	✓	✓	✓	8 S	16 Ss	16	8	16 S	–	–

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

middle ground, retaining subtle distinctions in local intensities while compressing overall scene brightness. Nonetheless, a fully static colour mapping with multi-resolution temperature partitioning remains only partially explored. Our proposed DTMRE technique complements this area by employing a fixed, discrete set of colour gradients in conjunction with multi-resolution segment interpolation, effectively ensuring reliable per-temperature encoding without losing the critical detail required for accurate object detection and classification.

2.3. Multi-resolution depth encoding

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

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

3. Multimodal hardware system

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

The sensor array configuration includes:

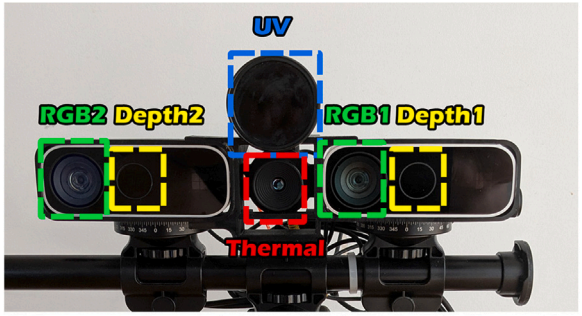


Fig. 1. Multimodal sensor array comprising 2×RGB, 2×Depth + NIR (850 nm), 1×LWIR (8–14 μm), and 1×UV (300–420 nm).

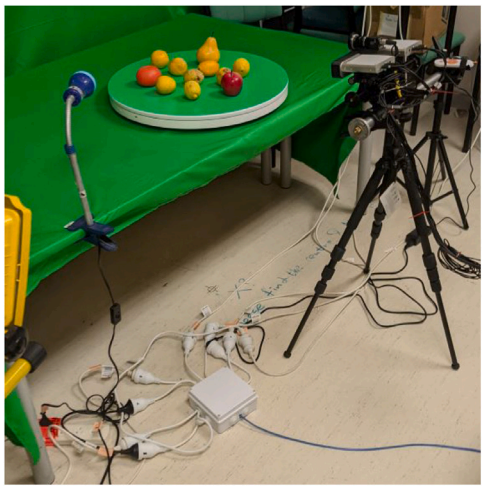


Fig. 2. Capturing setup in a controlled laboratory environment.

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

This multimodal setup forms the foundation for the MM5 dataset by ensuring accurate and consistent data capture across diverse spectral domains. By carefully aligning and calibrating each sensor, we establish a reliable framework for generating high-quality, multimodal training data, facilitating advancements in cross-spectral learning and sensor fusion research, as depicted in Fig. 1.

4. Capturing setup

We set up the data capture system in a controlled laboratory environment to ensure consistent and reproducible acquisition conditions across all modalities. Ambient lighting remained constant throughout all recording sessions to minimise external illumination variability. We utilised a green cloth backdrop for certain scenes to simplify background segmentation and enable subsequent replacement, thus facilitating data augmentation through background variation. However, the backdrop introduced unintended reflections of infrared (IR) illumination emitted by the Time-of-Flight (ToF) depth sensor, resulting in gaps within the depth data. These artefacts were deliberately retained in the dataset, presenting a realistic challenge. Furthermore, to maintain consistent exposure levels across captures, the RGB cameras were operated with fixed exposure settings, preventing automatic adjustments that could introduce inconsistencies in brightness and contrast. This setup allows for controlled acquisition of overexposed and underexposed scenes, ensuring a diverse dataset that reflects real-world lighting conditions. To further guarantee stability and precise alignment during data capture, the sensor array was securely mounted into a custom-designed, rigid 3D-printed frame, which was firmly affixed to a tripod (Fig. 1). We performed calibration of the thermal, UV, and right RGB cameras relative to the left RGB camera, thereby establishing precise spatial correspondence across all imaging modalities. The complete data acquisition setup is shown in Fig. 2.

4.1. Lighting

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

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

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

5. Capturing software

To develop a high-performance software solution for multimodal data acquisition, we implemented the system in C++ due to its efficiency, low latency, and robust hardware integration capabilities. The software interfaces with the Kinect SDK and the thermal imager SDK while simultaneously integrating the UV camera video stream and capturing metadata for each acquisition. This setup enables synchronised image capture across all modalities, continuous video stream recording, and incorporates a stereo calibration algorithm [61] conveniently integrated with OpenCV [62]. This calibration algorithm facilitates the estimation of the translation vector, rotation matrix, and distortion coefficients. We store the raw images and corresponding camera parameters, enabling alignment computations in a dedicated post-processing stage, which ensures flexibility for refining and optimising alignment. Additionally, to enhance automation and dataset diversity, the system incorporates a relay array for dynamically controlling scene illumination and a motorised turntable that rotates objects by 120 degrees, facilitating the acquisition of three distinct viewpoints per scene. Fig. 3 shows the capture and real-time alignment that allows for the testing and fine-tuning of the obtained camera parameters.

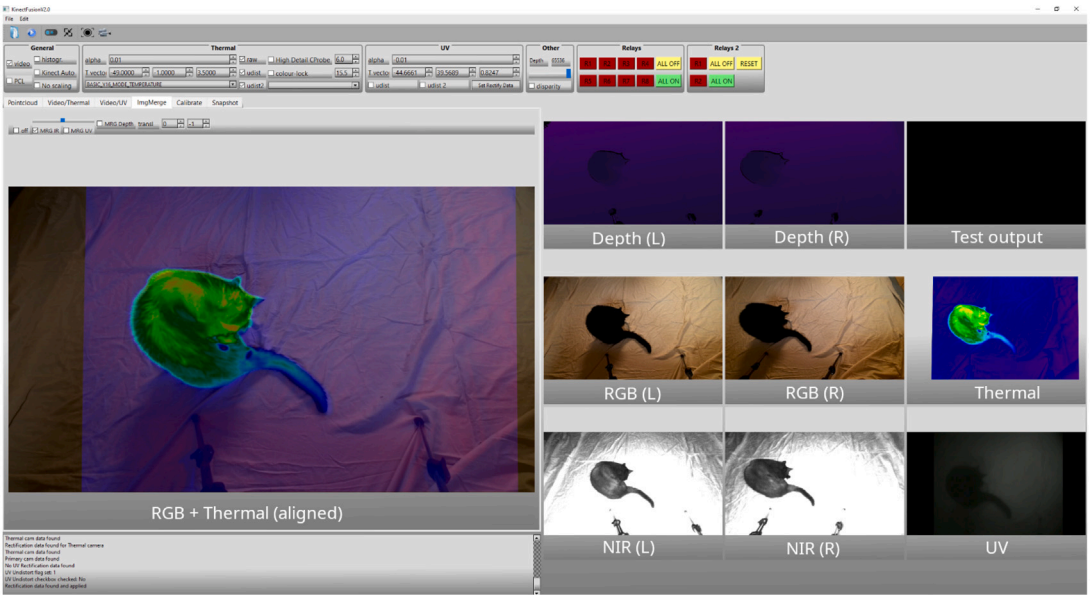


Fig. 3. Screenshot of the capturing software with real-time overlay and rectified thermal data showing a cat. The panel on the left shows the thermal data aligned with the left RGB data. The nine panels on the right, starting from top left to bottom right, show depth L+R, test output, RGB L+R, thermal, NIR L+R and UV.

6. Camera calibration and registration

For successful multimodal data level fusion using multiple modalities, it is crucial to acquire the data from these modalities correctly aligned. This can pose a challenge, as the sensors used for each modality have different fields of view (FOV), resolutions, and sensing capabilities. To facilitate data-level fusion, the system was calibrated by determining the intrinsic (pinhole camera model parameter matrix) and extrinsic (estimation of the relative sensor poses) parameters of each camera, which can then be used to align the data. Based on the pinhole camera model, this calibration has been simplified using a stereo calibration process [61], which can be applied using these and similar modalities. This method has been implemented in numerous studies in different ways, as summarised by Brenner et al. [1]. Fig. 5 shows the pattern matching using stereo calibration. However, uneven and fluctuating heat can complicate the calibration of the thermal camera. To achieve a more uniform heat distribution, the backside of the calibration board was covered with copper plates, as shown in Fig. 4, and a heating mat was placed over them. This approach helped to stabilise the temperature during calibration. Additionally, since the UV camera can detect wavelengths extending into the lower bounds of the visible spectrum (up to 420 nm), we removed the UV filter lens during the alignment process. This allowed us to use the captured grayscale image for alignment without requiring a dedicated setup for the UV modality. Furthermore, because stereo calibration requires uniform image resolutions, we applied lens distortion correction, scaling, and padding to the thermal and UV images prior to calibration. Fig. 5 shows an example set of RGB and thermal calibration images, including the calibration pattern matches generated by the stereo calibration algorithm.

After calibrating the intrinsic and extrinsic parameters of the RGB, thermal, and UV cameras, the thermal and UV images are aligned to the RGB coordinate system through a projection transformation. However, this alignment is optimised for a single reference plane in the scene; objects that lie substantially nearer or farther than this plane exhibit misalignment due to parallax effects. Consequently, the current approach ensures robust alignment in the primary plane of interest, with gradually increasing deviations as objects move away from that plane. Because the Kinect SDK automatically aligns the depth data to the RGB camera, we can directly utilise it without further modification. However, the intensity image, representing the 850 nm near-infrared

Fig. 4. Calibration board backside with partially applied copper plates.

(NIR) reflectance, is not a standard output of the Kinect SDK and requires additional processing. There are two primary approaches to obtaining this image. The first method extracts intensity values using the depth alignment process, ensuring direct correspondence with the depth map but limiting intensity information to pixels with valid depth data. The second method employs a synthetic flat depth image to bypass this limitation, allowing the retrieval of a complete intensity image that captures the full 850 nm NIR reflectance. However, as this image retains the field of view (FOV) of the depth sensor rather than the RGB camera, post-processing is required to correct FOV discrepancies. Additionally, the alignment process performed by the Kinect SDK operates discretely across depth intervals due to the differing sensor FOVs, employing a closed-source algorithm. Although this process helps mitigate geometric distortions, it also introduces occlusion artefacts such as depth shadows and missing regions. To address these limitations, depth information from the stereo setup can be fused, reducing alignment inconsistencies and improving overall depth map completeness. Additionally, we include the calibration image sets and corresponding calibration data as part of our dataset, ensuring transparency and reproducibility of the calibration process while enabling researchers to validate or develop alternative calibration methods.

7. Labelling and post-processing pipeline

The labelling and post-processing pipeline is critical in preparing the dataset for multimodal analysis. This process ensures that annotations are consistently applied across different imaging modalities

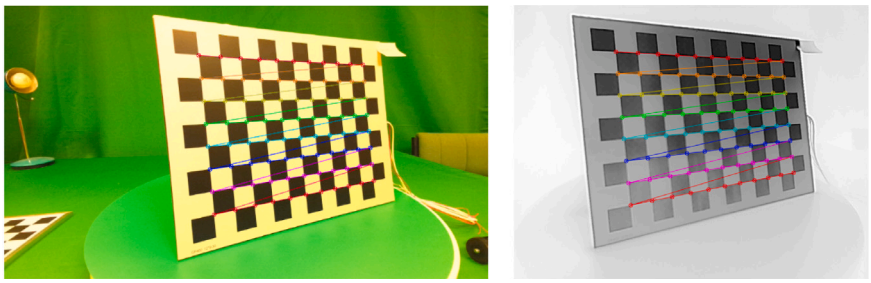


Fig. 5. Calibration images of RGB (left) and thermal (right) captured at approximately 30 cm from the sensor. The overlaid lines illustrate the pattern recognition process of the stereo calibration.

while maintaining spatial alignment between RGB, depth, thermal, and ultraviolet (UV) data. Given the inherent differences in sensor properties, including variations in field of view (FOV), resolution, and spectral characteristics, an efficient annotation workflow is necessary to achieve high-quality labelled data suitable for downstream tasks such as object detection, segmentation, and classification. For this, we employed a structured annotation workflow to facilitate accurate labelling, initially focusing on the RGB images. Once annotated, these labels were reprojected onto the thermal and UV images using transformation matrices derived from the camera calibration. This method ensures that the annotations are accurate and consistent across modalities, enabling cross-modal learning and sensor fusion. Following the automated labelling process, a series of post-processing steps are applied to refine the annotations and improve alignment. These steps include depth-based adjustments and annotation corrections to compensate for the different viewing angle. A detailed overview of the entire pipeline is shown in Fig. A.14 and sections describing the annotation framework in Label Studio, the label export process and the post-processing techniques applied to ensure high-quality multimodal alignment can be found in Appendix A while algorithm Alg. 1 below summarises the key steps in the multimodal image alignment process.

Algorithm 1 Multimodal Image Alignment

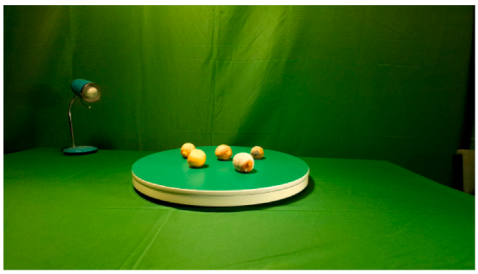
1: *Load camera calibration parameters* (including intrinsic and extrinsic matrices).
2: *Read raw images* from all modalities (RGB, Thermal, UV, Depth, and IR).
3: *Histogram Equalisation:* For Thermal and IR images, apply histogram equalisation to generate an 8-bit representation.
4: *Rectify images*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

7.1. MAR: Multimodal annotation remapping algorithm

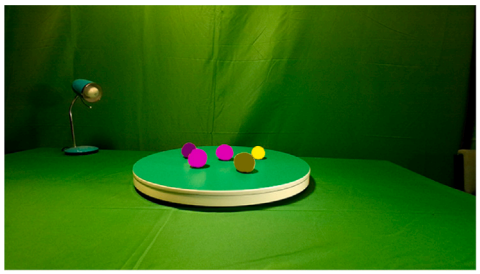
We generate labels for the unprocessed ultraviolet (UV) and thermal images to facilitate the exploration of fusion methods beyond early data-level fusion, which typically requires extensive calibration and can introduce latency in real-time processing. Our labelling approach, MAR (Multimodal Annotation Remapping), partially eliminates the labour-intensive task of manually labelling these modalities by reversing the same forward transformation that aligns thermal and UV images with the RGB labels. In essence, the RGB annotations, accurately aligned with the RGB images as shown in Fig. 6, are remapped onto the thermal and UV images(in their rectified state) and then distorted by reapplying lens distortion using the inverse of the original alignment transformation. This novel reverse mapping technique, to our knowledge, has not been reported previously in the literature, where multimodal fusion methods typically focus on early data-level fusion. In most cases, this approach yields accurate results, and even when minor discrepancies occur, it places the correct labels near their intended locations. This is particularly valuable given the inherent ambiguity in UV and thermal data, where insufficient contrast often makes it challenging to distinguish between classes clearly. Fig. 7 illustrates the target labels after processing, demonstrating how inverse camera matrices and transformation parameters transfer the labels from the RGB domain to the target modalities. In summary, MAR proceeds in four key steps: (1) inverse mapping (remapping annotations onto raw thermal/UV images), (2) re-distortion (re-applying lens distortion models), (3) depth-based refinement (adjusting labels based on average depth to account for FOV differences), and (4) edge-guided region growing (refining labels using a random walker [63] approach guided by Canny edges [64]). Although depth-based refinement, demonstrated in Fig. 8 using the Random Walker algorithm [63,65], is a key component of our approach to maximise alignment accuracy, it can be omitted when depth information is unavailable or insufficient. In such cases, MAR still performs reliably by relying on the inverse transformation and re-distortion steps, followed by edge-guided refinement. Additional pseudocode and implementation details are provided in Appendix D, where we sequentially describe each step: starting with inverse mapping and re-distortion (Appendix D.2), and concluding with depth-based correction (Appendix D.3) and edge-guided random walker refinement (Appendix D.4). A flowchart summarising the overall sequence of steps is provided in the Appendix (Fig. D.17). The MAR algorithm primarily serves as an automated label transfer and initialisation tool to reduce annotation effort. The remapped labels are intended for manual refinement, supported by subsequent edge-guided and machine learning-based segmentation techniques, to produce high-quality ground truth. Our qualitative visualisations in Figs. 7 and 8 demonstrate that MAR provides spatially consistent and accurate label placements across modalities, substantially accelerating the annotation process and improving consistency compared to manual labelling from scratch.

To quantitatively evaluate the accuracy of MAR-generated annotations, we compared them against the final manually corrected labels using mean Intersection over Union (IoU) and pixel accuracy metrics per class. The results, summarised in Table D.10 in the Appendix, demonstrate that MAR-generated labels closely match the manual corrections, achieving mean IoU and accuracy values generally between 80% and 90% for most classes. These findings confirm that MAR provides a reliable automatic initialisation that substantially reduces manual annotation effort and facilitates the generation of high-quality labels.

Overall, this novel approach, which combines inverse geometric transformation, depth-based FOV correction, and edge-guided region growing, constitutes a robust solution for efficiently transferring annotations from RGB to thermal and UV domains.

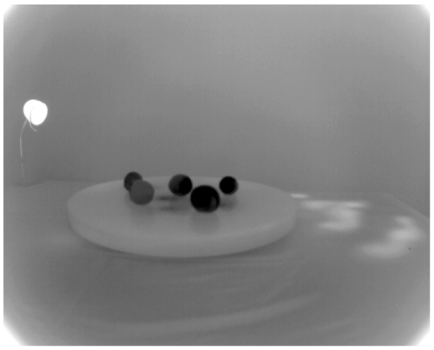


(a) RGB image without labels.

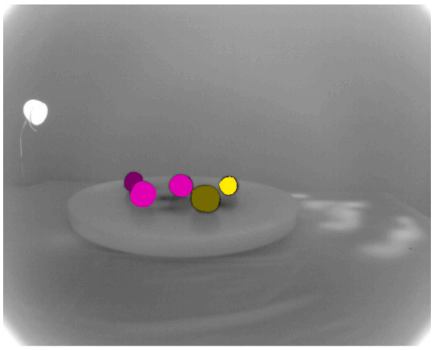


(b) RGB image with labels.

Fig. 6. RGB images showing colourised label overlay. The left image is without labels, and the right image includes manually created labels. The three pink-toned lemons on the left are good, while the two yellow-toned lemons indicate mold.

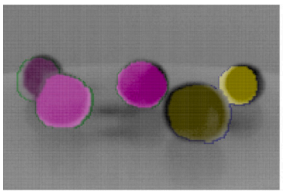


(a) Thermal image without labels.

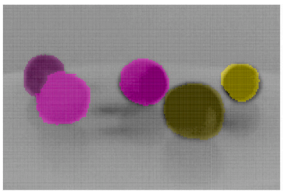


(b) Thermal image with labels.

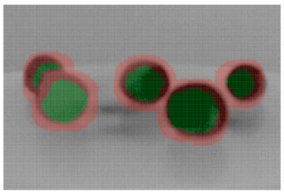
Fig. 7. Thermal images illustrating reprojection results, shown without labels (left) and with calculated labels (right).



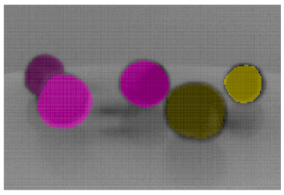
(a) Remapped and re-distorted labels



(b) Labels after depth correction.



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



(d) Labels after Random Walker.

Fig. 8. MAR refinement process. (a) Initial remapped and re-distorted labels. (b) Labels after depth-based correction. (c) The calculated seed (green) and growth (red) areas. (d) The final, refined labels after applying the Random Walker [63,65] optimisation.

7.2. Final image generation and storage

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

8. MM5 dataset

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

8.1. Structure

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

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

An example filename for an RGB image with light setting 5 is: `1_5_20240716_130310_143_rgb.png`.

The captured raw data and the transformation outputs provided by the Kinect SDK are available for the depth and IR modalities. These files are differentiated by a postfix: `_raw` for raw data and `_tr` for transformed data. Both depth and IR images are stored as 16-bit single-channel images. For the thermal (LWIR) modality, multiple representations are provided:

- Raw 16-bit images (`_lwir16`)

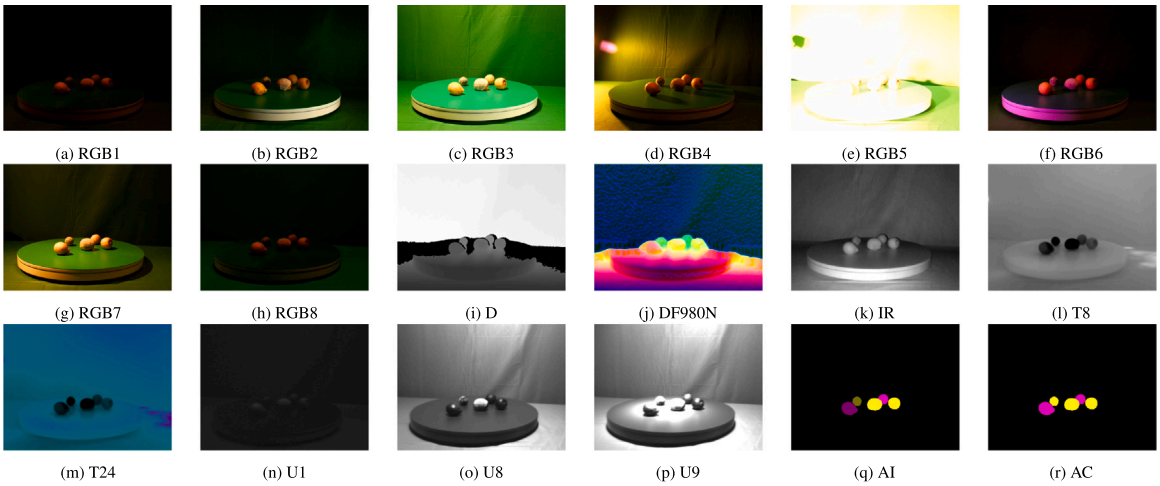


Fig. 9. Set of images from the dataset.

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

The encoded LWIR images are included for convenience, as they can be derived from the raw data.

The aligned and cropped dataset is generated by selecting a subset from the raw data, since the raw data includes additional unlabelled images, and renaming these selected images sequentially, starting from 1, to ensure consistent filenames across all modalities captured simultaneously. The processed data are organised into the following folders:

- `ANNO_CLASS`
- `ANNO_INST`
- `ANNO_VIS_CLASS`
- `ANNO_VIS_INST`
- `D`
- `D_Focus`
- `D_Focus960N`
- `D16`
- `I`
- `I16`
- `META`
- `RGB1`
- `RGB2`
- `RGB3`
- `RGB4`
- `RGB5`
- `RGB6`
- `RGB7`
- `RGB8`
- `T8`
- `T16`
- `T24`
- `U1`
- `U8`
- `U9`

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

8.2. Images

Fig. 9 displays the same scene captured under eight different light settings, as described in Section 4.1. The figure includes eight RGB images (RGB1 through RGB8) that illustrate the effects of varying illumination alongside a depth image (D), a processed depth image (DF980N) and an infrared image (IR). It also shows two thermal encodings (T8 and T24) and three UV images (U1, U8, and U9). Additionally, the annotation images, for object instance (AI) and class (AC) labels, are provided to demonstrate the corresponding ground truth. These images offer a comprehensive view of the scene and underscore the diverse conditions captured in the MM5 dataset.

8.3. Thermal raw data

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

$$T_{\text{Celsius}} = \frac{T_{\text{raw}}}{64} - 273.15, \tag{1}$$

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

8.4. Thermal 8-bit data

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

8.5. Thermal 24-bit data

The thermal 24-bit data in the MM5 dataset is produced by applying our novel DTMRE algorithm for static colour mapping to the raw temperature values. This mapping utilises a predefined gradient of distinct colours, each corresponding to specific temperature intervals. For further details, please refer to Section 9.1.

8.6. Depth raw data

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

8.7. Depth 8-bit data

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

8.8. Depth focused 8-bit data

This variant of the depth image is produced by applying our novel ADMRE algorithm for the adaptive compression strategy detailed in Section 9.2, resulting in an 8-bit representation that preserves fine detail where depth variation is most pronounced. Non-essential regions are compressed at lower resolution, reducing noise and enhancing focus on critical structures. Consequently, the final 8-bit output provides a compact yet detailed view of salient depth information.

8.9. Depth focused 24-bit data

The same ADMRE 9.2 algorithm as for the 8-bit focused data is used, but with an additional step that packs the data in a 24-bit format. The first two channels store the quantised depth values, while the third channel encodes the computed surface normals, allowing a visual representation of both geometry and spatial orientation. This approach offers a more complete scene depiction, combining depth-focused compression with local angular detail for downstream tasks.

8.10. Meta data

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

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

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

Table 2
Label distribution table.

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

8.11. Labels

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

Figs. 9(q) and 9(r) illustrate these two approaches. The former, AC (Annotation Class), displays the class labels, while the latter, AI (Annotation Instance), shows the object instance labels. This dual annotation strategy conforms to standard practices in semantic and instance segmentation that are exemplified by widely used datasets such as PASCAL VOC [66], MS COCO [67], and NYU [23], as well as by segmentation methods such as Mask R-CNN [68]. This approach facilitates both class-level analysis and object-level detection.

8.12. Classes

Currently, the dataset comprises 14 top-level classes and a total of 32 labelled classes across 324 scenes, as detailed in Table 2. Ongoing efforts will continue to expand the dataset by incorporating additional scenes, videos, and annotated classes in future releases.

The chart in Fig. 10 illustrates the total class distribution, highlighting the relative frequency of each class variant within the total frames that encompass the class.

8.13. Calibration data

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

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

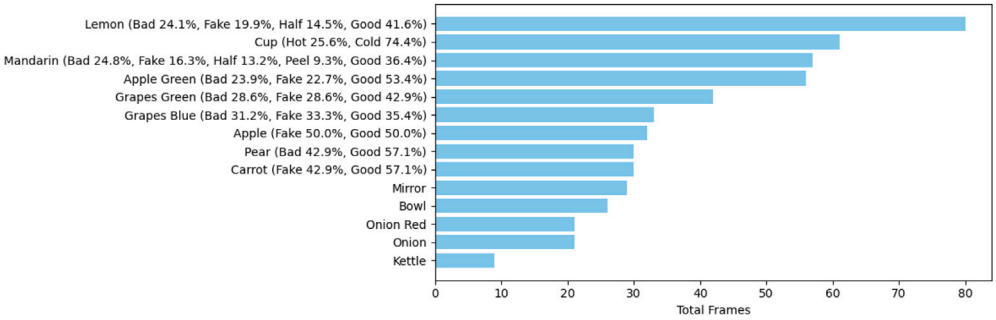
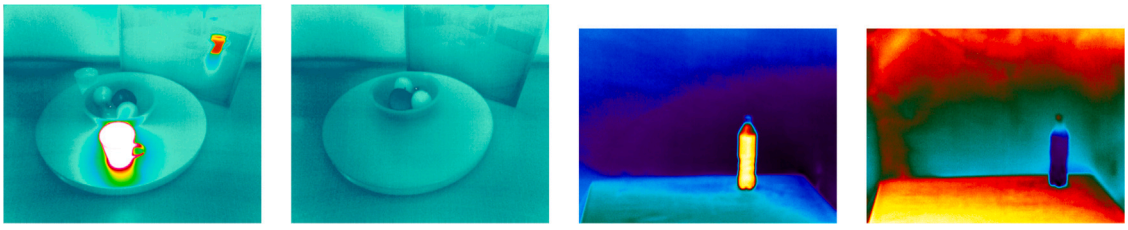


Fig. 10. Distribution of unique label occurrences. Percentages shown in parentheses represent the frequency of subclass occurrences (e.g., good, bad, fake) among all class occurrences. Due to mixed-scene acquisitions, subclasses may co-occur together within a single frame.



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

Fig. 11. Comparison of 24-bit thermal image representations. Images (a) and (b) show MM5's DTMRE method, while images (c) and (d) depict images from the VDT2048 dataset processed with automatic gain control (AGC) and a colour scheme. Images (a) and (c) show hot objects, whereas images (b) and (d) represent cold objects.

Table 3
Defined primary colour transitions used in our DTMRE demonstration.

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

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

9. Data pre-processing

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

9.1. Processing thermal data with DTMRE

The proposed 24-bit thermal data representation in the MM5 dataset applies a *static* colour mapping to raw temperature values, thereby addressing long-standing issues of inconsistent contrast that arise in dynamically adjusted schemes [38,39]. Specifically, a predefined gradient is segmented into multiple intervals and populated via linear interpolation at each step. This gradient, spanning 2455 distinct colours, transitions through a series of perceptually distinct hues (e.g. black to teal, teal to purple), ensuring that temperature distinctions remain visually salient, resulting in a refined palette that conveys subtle temperature cues. The raw thermal readings are first translated into degrees Celsius by applying Eq. (1). The temperature scale is then divided into segments of varying resolution: a fine-grained *core* interval (for example, 14 °C to 30 °C), surrounded by broader segments for higher or lower temperatures (e.g. 30 °C to 40 °C, 40 °C to 100 °C, and 0 °C to 14 °C). A piecewise linear function determines the appropriate index for a given temperature, ensuring that out-of-bounds values are clamped. By returning the colour at the computed index, this scheme produces a stable 24-bit representation that faithfully captures pixel-wise thermal variation. Table 3 outlines the primary colour transitions used by this method.

This static mapping method overcomes the shortcomings of Automatic Gain Control (AGC) algorithms, which convert raw 16-bit data (0–65,535) into 8-bit images (0–255), compressing the data and reducing detail. Although AGC algorithms enhance contrast and brightness to emphasise contextual details [69], their dynamic adjustments can lead to inconsistent representations between frames. In contrast, our novel DTMRE method preserves environmental features regardless of object temperature. Fig. 11 shows two sets of thermal images: images a and c depict a hot object, while images b and d represent cold objects. In the

dynamic range images taken from the VDT2048 dataset [33] (images c and d), noticeable variations in the appearance of the background and table are observed despite no actual changes, whereas our DTMRE processing maintains a stable visual representation. Even with a cup of boiling water in the frame, the minute temperature differences of the fruit in the bowl are visible. This consistency is crucial for the reliable processing of thermal data in form of visual information. It should be noted that slight variations in the thermal output may occur over time and during non-uniformity correction (NUC) procedures for uncooled thermal cameras, potentially introducing minor discrepancies in the measured values and, consequently, in the visual appearance of the images. The consistency provided by the DTMRE encoding is crucial when detecting minute temperature differences, such as those between rotten and good fruits.

The complete pseudocode for the DTMRE algorithm is provided in Appendix B.

9.1.1. Evaluation

To evaluate the performance of the T8 (normalised 8-bit thermal image 8.4) and T24 (24-bit DTMRE processed thermal image) representations, we process the full 324 scenes from the MM5 dataset. The training and evaluation images are defined in the files list_train_f.txt and list_eval_f.txt, respectively. We utilised a CMX model [70], a SegFormer-based segmentation network, and only used low-light (RGB1) images as supplementary input. This design ensures that thermal data remains the dominant source of information for segmentation. We have trained each model from scratch for 500 epochs on the MiT-B0 backbone, and the final segmentation metrics are summarised in Table 4. Detailed class-level results can be found in Table B.7.

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

The quantitative evaluation, summarised in Table 4 and detailed in Table B.7, demonstrates that the proposed 24-bit DTMRE thermal representation (T24) consistently outperforms both the normalised 8-bit thermal image (T8) and the established contrast enhancement methods PHE, CLAHE, and MSR. Specifically, T24 achieves the highest mean Intersection over Union (79.08%) and mean pixel accuracy (86.67%), surpassing MSR and T8, which attain a mean IoU of 72.67% and 72.29% and a mean pixel accuracy of 82.93% and 81.94% respectively. Among the traditional methods, Multi-Scale Retinex (MSR) performs best with a mean IoU of 72.67%, followed by CLAHE (69.12%) and PHE (60.52%). Notably, T24 provides marked improvements in challenging classes such as *Mandarin Peel* and *Pear Bad*, indicating its superior ability to preserve thermal details critical for segmentation accuracy. These results highlight the effectiveness of the DTMRE encoding in enhancing thermal image quality for robust multimodal fusion, particularly under low-light conditions where thermal cues are essential.

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

Table 4
Performance of Thermal Image Preprocessing. A comparison of our proposed DTMRE method (T24) against established baseline algorithms (PHE, CLAHE, MSR) and our normalised (T8) image. Metrics include Intersection over Union (IoU), Pixel Accuracy (Acc), and Mean Rank, with IoU and Acc reported as percentages (%).

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

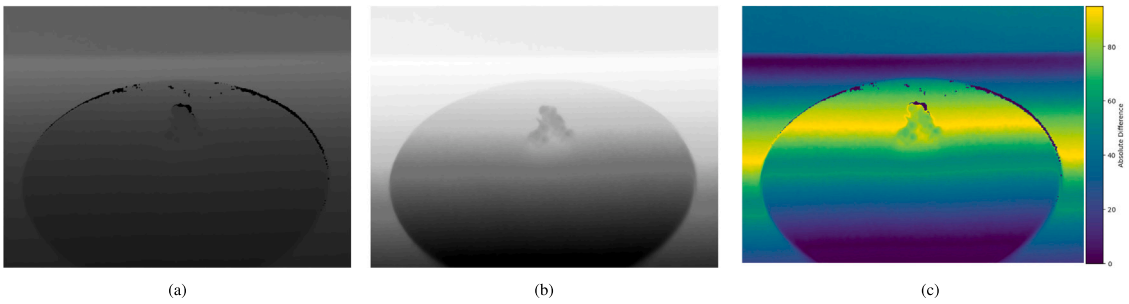
Table 5
Comparison of ADMRE processed (P), normalised (N), equalised (E), and raw (R) intensity data in the indicated green (G) and red (R) area of the images in Fig. 12(c) and (d).

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

9.2. Processing depth data with ADMRE

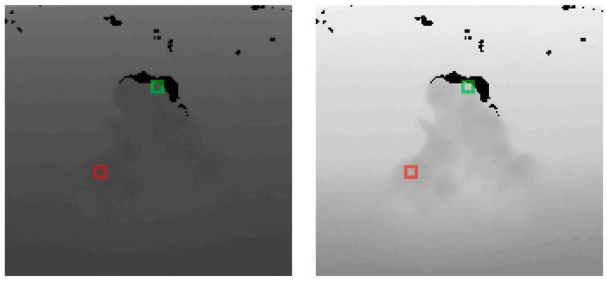
In the MM5 dataset, depth information is captured using a Time-of-Flight (ToF) sensor, which intrinsically provides high-fidelity range measurements with millimetre resolution. Over a distance of approximately 3 m, these sensors generate up to 3000 discrete depth values, increasing further for longer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

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



(a) (b) (c)

Fig. 12. (a) Normalised 8-bit image, (b) focused 8-bit image, and (c) Pixel-wise absolute difference between ADMRE processed and normalised image.



(a) (b)

Fig. 13. (a) Zoomed-in normalised 8-bit image, and (b) zoomed-in focused 8-bit image of grapes.

Table 6
Overall mean IoU (%) for each depth encoding method combined with RGB3, along with summary metrics and estimated processing time per image (ms) for 640 × 480 resolution. DF980N achieves both superior segmentation accuracy and significantly lower computational cost compared to HHA.

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

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

The complete pseudocode for the ADMRE algorithm is provided in Appendix C.

9.2.1. Evaluation

Fig. 12 compares normalised 12(a) and 8-bit ADMRE-processed 12(b) depth representations of an example scene. A detailed close-up of an object of interest (grapes) in this scene is shown in Figs. 13(a) and 13(b), highlighting two specific regions (marked in green and red) selected for pixel intensity measurement. In addition to these processed representations, intensity values were recorded from equalised and raw depth data as well. These measurements are summarised in Table 5. Furthermore, Fig. 12(c) visualises the absolute pixel-value differences between the normalised and ADMRE-processed images, emphasising pronounced deviations within the detected peak region.

The results demonstrate a value difference of 33 in the raw data, corresponding to a distance of 33 mm. In the ADMRE processed image, there are 30 values for this region, whereas in the normalised image, we obtained only 14 values. Additionally, the histogram's equalisation yielded a mere seven value difference. These effects amplify with the distances present in a frame, as a more pronounced initial value range results in a greater compression of values during the normalisation or equalisation process. We evaluated the effectiveness of our approach using CMX with a Segformer-B0 backbone, a state-of-the-art vision transformer framework for segmentation, trained from scratch on 324 scenes of the MM5 dataset for 250 epochs. The training and evaluation images are defined in the files `list_train_f.txt` and `list_eval_f.txt`, respectively. Table 6 summarises the mean IoU per method (%) for the three variants: standard depth (**D**), depth focus (**DF**), and depth focus with 980 discrete levels plus normals (**DF980N**) as well as HHA encoded depth. The full class-level results can be found in the Appendix in Table C.9. Since the network only has access to visual and geometric cues, discriminating among defective (*bad*), plastic (*fake*), and (*good*) produce classes poses a significant challenge due to the similarity of the objects in shape and colour.

Despite these inherent difficulties, the results in Table 6 indicate that **DF** and **DF980N** outperform the baseline **D** in terms of mean IoU, frequency-weighted IoU, and pixel-level accuracy. Importantly, **DF** and **DF980N** can be computed in approximately 25 ms per image, whereas HHA encoding is substantially slower, requiring over 500 ms per 640 × 480 image with the original method and still around 260 ms even with a recent optimised implementation [77]. We attribute the observed segmentation improvements to the selective preservation of fine-grained depth details in salient image regions. In contrast, normalised 8-bit depth often fails to capture subtle curvature and spatial variation crucial for distinguishing nuanced object classes [53]. By dedicating higher resolution to areas with strong depth gradients while simultaneously allowing coarser encoding in low-variation zones, our approach helps minimise quantisation artefacts like contour banding or edge distortion that degrade downstream performance [55,60]. These findings align with previous research on region-based compression and saliency-driven encoding [49], confirming that refined geometric features can significantly enhance scene segmentation [57]. More importantly, they demonstrate that curated depth encoding can serve as a decisive factor in differentiating between visually similar categories, thus validating the key design goals of our proposed method.

10. Challenges and future work

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

- **Multimodal Sensor Calibration and Alignment:** Although comprehensive calibration procedures have been implemented to align RGB, depth, thermal, ultraviolet (UV), and near-infrared (NIR) sensors, misalignments can still occur due to lens distortions, differences in fields of view (FoV), and parallax effects. More robust calibration and alignment methods, potentially aided by deep learning, could further mitigate alignment errors, compensate for parallax-induced discrepancies, and ensure consistent feature fusion across modalities.
- **Reflective Surfaces and Missing Data:** Scenes containing mirrors, metallic objects, or liquids can result in regions of missing or invalid depth data due to reflections or sensor interference. Although these scenarios are realistic and highlight the importance of robust data preprocessing, the resulting missing or noisy depth data remains a significant challenge. Future studies could focus on advanced hole-filling strategies or deep learning-based inpainting methods that retain cues about reflective surfaces.
- **Thermal Consistency Under Varying Conditions:** The 24-bit static thermal mapping in MM5 counters the dynamic nature of AGC algorithms, but uncooled cameras can still exhibit measurement fluctuations and drifts, especially during extended operation or pronounced environmental changes. Non-uniformity correction (NUC) processes may also introduce subtle shifts or discontinuities over time. Future work could investigate techniques to minimise these effects, enabling stable thermal representations and ensuring reliable pixel-wise temperature measurements under variable conditions.
- **Large-Scale Data and Real-Time Fusion:** Although MM5 includes diverse objects, lighting conditions, and sensor modalities, further expanding the dataset with extensive indoor and outdoor scenes would enhance its benchmarking capabilities. Many real-time applications, such as robotics and AR/VR, require efficient on-device multimodal data fusion. Future studies could focus on developing fusion pipelines optimised for real-time processing in resource-constrained environments.
- **Annotated Multimodal Datasets with Complex Tasks:** The current version of MM5 supplies aligned and unaligned annotations suited to segmentation, object detection, and classification tasks. However, more advanced challenges remain. For instance, spatio-temporal action recognition would require temporal annotations and sequences of frames across modalities. At the same time, 3D reconstruction and material classification would require correspondingly richer labels detailing structure and surface properties. Extending the dataset with these additional annotations and establishing benchmarks that span multiple tasks would enable researchers to investigate modality-specific effects more thoroughly and advance the development of robust fusion techniques that leverage the complementary strengths of each sensor modality.

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

11. Conclusion

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

CRediT authorship contribution statement

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

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. Labelling process

A.1. Labelling process overview

See Fig. A.14.

A.2. Labelling and post-processing: Label studio

To facilitate the annotation process, we utilised *Label Studio*, an open-source data labelling platform designed for multimodal annotation tasks [78]. Label Studio provides a flexible web-based interface that supports various annotation types, including image segmentation, classification, and object detection. The platform's ability to handle custom labelling workflows made it well suited for our multimodal dataset, where labels initially created for the RGB images are later reprojected onto the thermal and UV modalities. One of the primary advantages of using Label Studio is its compatibility with deep learning-assisted annotation. To enhance the efficiency and accuracy of the annotation process, we integrated Meta's *Segment Anything Model* (SAM) [79] to help label RGB images. SAM is a powerful image segmentation model that enables automatic region selection, significantly reducing the manual effort required to annotate complex scenes. By leveraging SAM within Label Studio, we streamlined the annotation pipeline, improving consistency and minimising human-induced errors. In a first step, RGB images were set up for annotation, and later projects were configured for the thermal and UV modalities by copying the RGB configuration. When the RGB labels are reprojected onto the thermal and UV images using transformation matrices obtained from the multimodal calibration process, the resulting labels are created in those target

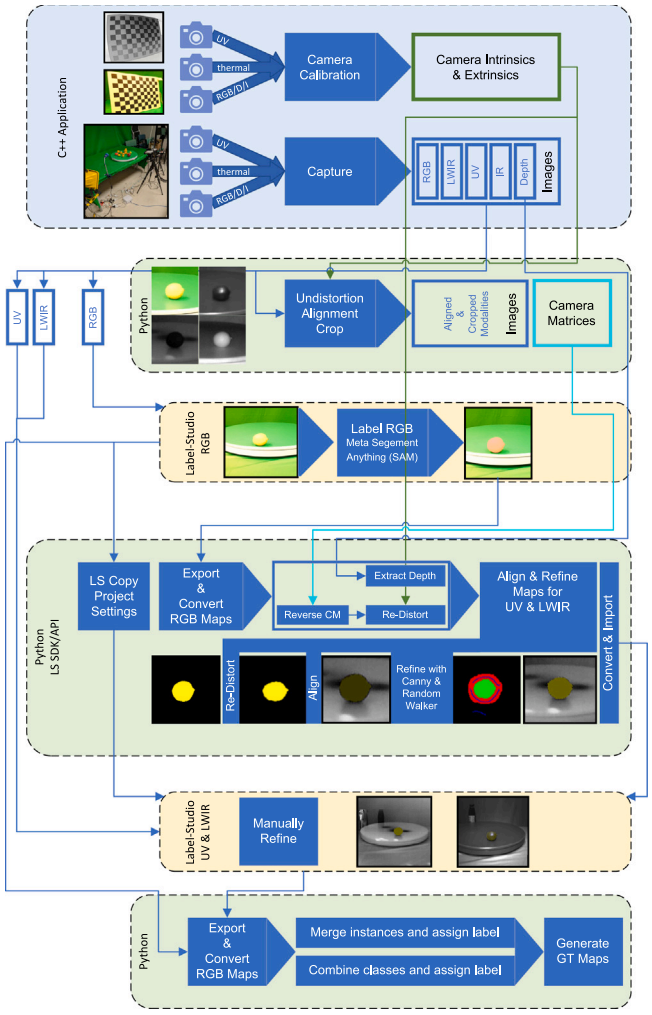


Fig. A.14. Labelling process pipeline.

projects. Label Studio provided several key benefits in the annotation and dataset management workflow for MM5: it offered a web-based interface that enabled efficient collaborative annotation across multiple modalities, facilitating streamlined data labelling; it supported deep learning models by integrating with segmentation models such as SAM to pre-label data, significantly reducing manual annotation effort; it allowed multimodal compatibility for handling RGB, thermal, and UV images within a unified framework, ensuring consistency of annotation across different sensor outputs; it featured export options in JSON and COCO formats, facilitating seamless integration into the downstream processing pipeline; and it offered a Python SDK [80] integration that supported automated project setup, replicate configurations, and extracted metadata, with export/import functionality allowing calculated UV and thermal labels to be re-imported for iterative refinement. By incorporating Label Studio into our annotation workflow, we established a scalable and efficient framework to generate a high-quality labelled dataset across multiple imaging modalities, ensuring that the labelled data remain consistent, well organised, and aligned for subsequent multimodal analysis.

A.3. Multimodal image alignment and processing

Aligning multimodal images involves rectifying and registering data from different sensors to ensure spatial consistency across all modalities. This section describes the pipeline used to generate accurately aligned images from the RGB, depth, thermal, and ultraviolet (UV) captures by applying camera calibration parameters and the re-projection of labels onto the original, distorted thermal and UV modalities. The process begins with loading the camera calibration parameters, where the intrinsic and extrinsic parameters of each sensor are retrieved from the calibration files. Next, the RGB annotations on class and instance levels are exported from Label Studio. Subsequently, the captured RGB, thermal, UV, depth, and infrared (IR) images are read into the processing pipeline. Distortion correction is then applied to thermal and UV images using the camera's intrinsic parameters, after which thermal and UV images are aligned to the RGB coordinate system using projection transformation matrices. Additional postprocessing transformations are applied to correct minor misalignments caused by field-of-view differences based on the depth of labelled objects. Finally, the pipeline produces cropped and full-resolution aligned outputs for further processing. A flow diagram summarising these steps is shown in Fig. A.14, providing an overview of the entire multimodal image alignment process. The following subsections detail the methods used for Image Rectification and Alignment, the Processing Pipeline Algorithm, and the MAR: Multimodal Annotation Remapping Algorithm.

A.3.1. Image rectification and alignment

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

Appendix B. DTMRE implementation

B.1. Detailed evaluation results

See Table B.7.

B.2. DTMRE algorithm

See Algorithm 2.

Appendix C. ADMRE implementation

C.1. Processing steps

Each detected peak is characterised by a width parameter, indicating the extent of significant depth variation, as demonstrated by

Algorithm 2 DTMRE: Deterministic Thermal Multi-Resolution Encoding

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

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

the KDE-based peak detection in Fig. C.15. Regions identified within these peak ranges are compressed using a finer resolution, whereas out-of-focus (OOF) areas, defined as regions closer or further than a specified focal window, are quantised more coarsely. Similarly, intervals between the identified peaks and the OOF regions ("gaps") undergo compression at a reduced resolution. This hierarchical approach prioritises depth detail around the most salient portions of the scene. Table C.8 details the region detection outcomes corresponding to Fig. C.15, applying compression factors of 10 for outer regions (depth values below 500 and above 900), 3 for the gap regions, and 1 for the peak regions. The object of interest, in this case, partially rotten green grapes visible in the corresponding RGB image (Fig. C.16), is centred around peak_712. By merging these individually compressed segments into a unified depth map and subsequently normalising the result, the algorithm effectively preserves critical spatial detail without incurring the computational overhead associated with uniformly high-resolution representations [55]. For visual comparison between standard normalisation and the ADMRE-processed result, Fig. 13 shows a close-up of the grapes.

C.2. Detailed evaluation results

See Table C.9.

C.3. ADMRE algorithm

See Algorithm 3.

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

Table B.7
Per-class mean Intersection over Union (IoU, %) for thermal preprocessing methods including Plateau Histogram Equalisation (PHE), Contrast Limited Adaptive Histogram Equalisation (CLAHE), Multi-Scale Retinex (MSR), MM5 normalised 8-bit thermal images (T8), and the proposed 24-bit DTMRE processed thermal images (T24), evaluated in combination with RGB1 dim light images. In addition to per-class IoU values, the table also reports mean pixel accuracy and the mean rank of each method across all classes, providing a comprehensive comparison of segmentation performance and ranking consistency over the full MM5 evaluation dataset.

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

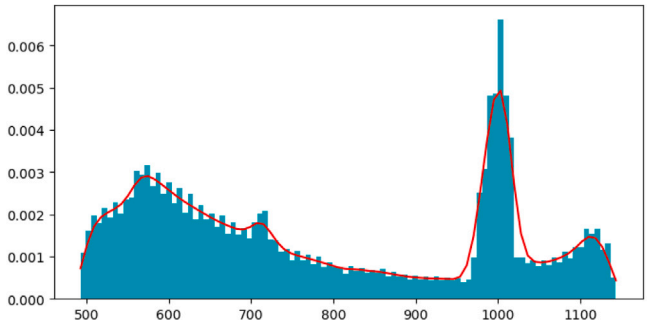


Fig. C.15. Raw data histogram, X axis with depth in mm and Y axis the pixel distribution with the KDE peak detection plotted in red.

Table C.8
Segmentation showing start/end indices and the number of unique values.

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

Table C.9
Per-class mean IoU (%) for each depth encoding method combined with RGB3, along with summary metrics and estimated processing time per image (ms) for 640 × 480 resolution. DF980N achieves both superior segmentation accuracy and significantly lower computational cost compared to HHA.

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

Appendix D. MAR implementation

D.1. Inverse mapping

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

Fig. C.16. RGB image of partially rotten green grapes.

$(m_x(x, y), m_y(x, y))$, where:

$$x_n = \frac{x - c_x}{f_x}, \quad y_n = \frac{y - c_y}{f_y}, \tag{D.1}$$

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

$$m_x(x, y) \approx u \quad \text{and} \quad m_y(x, y) \approx v, \tag{D.2}$$

then:

$$i_x(u, v) = x \quad \text{and} \quad i_y(u, v) = y. \tag{D.3}$$

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

$$i_x\big(m_x(x, y), m_y(x, y)\big) = x, \quad i_y\big(m_x(x, y), m_y(x, y)\big) = y, \tag{D.4}$$

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

D.2. Re-distortion

The re-distortion is implemented as follows: initially, pixel coordinates (x, y) are converted into normalised image coordinates (x_n, y_n) using the camera's intrinsic parameters. This is identical to the inverse mapping, since the mapping process also includes a distortion correction. Therefore, Eq. (D.1) is applied, where f_x and f_y are the focal lengths, and (c_x, c_y) is the principal point. The radial distance is then computed by

$$r^2 = x_n^2 + y_n^2. \tag{D.5}$$

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

$$x_d = x_n \left(1 + \alpha_x k_1 r^3 + \alpha_x k_2 r^4\right), \tag{D.6}$$

$$y_d = y_n \left(1 + \alpha_y k_1 r^3 + \alpha_y k_2 r^4\right), \tag{D.7}$$

Algorithm 3 ADMRE: Adaptive Depth Multi-Resolution Encoding

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

where α_x and α_y (`x_factor` and `y_factor`) are manually determined scale factors for axis-specific distortion limits, empirically set through experimentation, and k_1 and k_2 are the distortion coefficients derived from camera calibration.

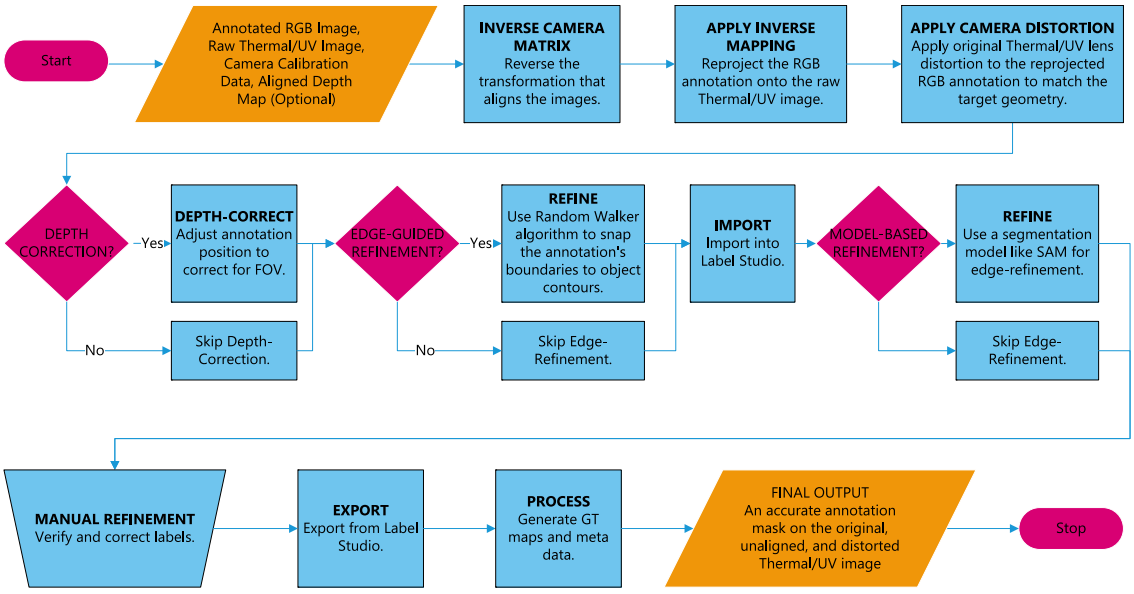


Fig. D.17. MAR flowchart.

Table D.10
Comparison of generated MAR labels against manually corrected thermal and UV labels (Mean IoU and Pixel Accuracy, %).

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

The distorted normalised coordinates are then converted back to pixel coordinates:

$$x_{\text{final}} = x_d\, f_x + c_x, \quad y_{\text{final}} = y_d\, f_y + c_y. \tag{D.8}$$

This mapping is applied across the image to produce the re-distorted image using the interpolation function `cv2.remap`.

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

D.3. Depth correction

Depth information is incorporated to refine the remapping. For each mapped object, the average depth is computed from the corresponding aligned depth map. This average depth value is then used to calculate correction factors that account for differences in the field of view (FOV) between the sensors, placing the labels more accurately in the target modalities. Figs. 8(a) and 8(b) provide a visual comparison of the annotations before and after applying the depth-based correction, demonstrating the enhanced alignment accuracy achieved by this approach.

D.4. Edge-guided random walker refinement

Finally, we optimise the labels by applying a random walker segmentation by first detecting edges using the Canny algorithm [64]. The detected edges are added to the seed area, and the initial RGB annotation is dilated by a factor (derived from the annotation map size) to define a maximum growth area. Subsequently, the annotation is eroded to create a consistent seed map. A Random Walker [63,65] algorithm expands each seed region based on pixel characteristics specific to the target modality, more accurately delineating object boundaries. Fig. 8(c) illustrates these seed and growth regions while Fig. 8(d) shows the result. For scenes containing multiple objects, we first identify all object seed regions, excluding overlapping or occluded areas from one another's growth domain to avoid growing into other objects in low contrast scenarios. Additionally, we compute the average pixel intensity within both seed and growth areas to dynamically adjust random walker parameters, improving results across varied contrast conditions. This method performs well in scenes with sufficient contrast in the thermal or UV data, though its effectiveness may be limited

Algorithm 4 MAR: Multimodal Annotation Remapping

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

in very low contrast conditions. Despite the geometric transformations and optional depth-based adjustments described, exact label alignment remains challenging when objects exhibit overlapping, complex geometries or when cameras capture significantly different viewing angles. Consequently, refinement methods such as random-walker-driven segmentation or advanced machine-learning techniques like the Segment Anything Model (SAM) [79] become essential for generating sufficiently accurate automated annotations. In our workflow, we integrated the SAM refinements within Label Studio using its SDK [80] and python, enabling us to automatically compute bounding boxes for each object and invoke SAM on a backend server to dynamically produce auto-annotations based on size, position, and class name. Although this approach can be adopted without Label Studio, the platform's flexibility and support for external models facilitated efficient iterative improvements to our automated labelling pipeline. In scenarios where SAM or similar segmentation models are employed, the random walker step may be omitted.

D.5. Flowchart

See Fig. D.17.

D.6. MAR algorithm

See Algorithm 4.

Data availability

The MM5 dataset introduced in this paper is publicly available under the following Figshare links: (1) Raw data https://figshare.com/ndownloader/files/56868443, (2) Aligned and cropped data https://figshare.com/ndownloader/files/55555457, (3) Label Studio annotations https://figshare.com/ndownloader/files/55555424, and (4) Calibration images https://figshare.com/ndownloader/files/55555421.

Additional resources, code examples, and updates are available via the project repository: https://github.com/martinbrennernz/MM5-Dataset. If you use this dataset in your research, please cite both this paper and the dataset DOI [12], for example: M. Brenner, N. Reyes, T. Susnjak, and A. Barczak (2025). MM5: Multimodal Image Dataset. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.28722164.

References

[1] 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.

[2] J. Muhovič, J. Perš, Joint calibration of a multimodal sensor system for autonomous vehicles, Sensors 23 (12) (2023) http://dx.doi.org/10.3390/s23125676, URL https://www.mdpi.com/1424-8220/23/12/5676.

[3] C. Li, D. Song, R. Tong, M. Tang, Illumination-aware faster R-CNN for robust multispectral pedestrian detection, Pattern Recognit. 85 (2019) 161–171, http://dx.doi.org/10.1016/j.patcog.2018.08.030.

[4] M. Abdul-Al, G. Kumi Kyeremeh, R. Qahwaji, N.T. Ali, R.A. Abd-Alhameed, The Evolution of Biometric Authentication: A Deep Dive Into Multi-Modal Facial Recognition: A Review Case Study, IEEE Access 12 (2024) 179010–179038, http://dx.doi.org/10.1109/ACCESS.2024.3486552.

[5] 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. Mechatronics 28 (3) (2022) 1558–1569.

[6] H. Wen, K. Song, L. Huang, H. Wang, J. Wang, Y. Yan, Hierarchical two-stage modal fusion for triple-modality salient object detection, Meas. 218 (2023) 113180.

[7] 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, http://dx.doi.org/10.1016/j.jksuci.2023.101702.

[8] 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.

[9] 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, http://dx.doi.org/10.1109/TIP.2024.3393365.

[10] J. Qiu, C. Jiang, H. Wang, ETFormer: An Efficient Transformer Based on Multimodal Hybrid Fusion and Representation Learning for RGB-D-T Salient Object Detection, IEEE Signal Process. Lett. 31 (2024) 2928–2932, http://dx.doi.org/10.1109/LSP.2024.3465351.

[11] N. Huang, Y. Yang, R. Xi, Q. Zhang, J. Han, J. Huang, Salient Object Detection From Arbitrary Modalities, 2024, arXiv preprint arXiv:2405.03352 Under review.

[12] M. Brenner, N. Reyes, T. Susnjak, A. Barczak, MM5: Multimodal image dataset, 2025, http://dx.doi.org/10.6084/m9.figshare.28722164, Dataset.

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

[14] 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.

[15] V.V. Kniaz, V.A. Knyaz, J. Hladůvka, W.G. Kropatsch, V. Mizginov, ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset, in: Computer Vision – ECCV 2018 Workshops, in: Lecture Notes in Computer Science, 11134, Springer, Cham, 2019, pp. 606–624, http://dx.doi.org/10.1007/978-3-030-11024-6_46.

[16] Y. Shi, Y. Li, S. Liang, H. Chen, Q. Miao, MGR-Dark: A large multimodal video dataset and RGB-IR benchmark for gesture recognition in darkness, in: Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM), Association for Computing Machinery, 2024, pp. 2321–2330, http://dx.doi.org/10.1145/3664647.3681267.

[17] D. Guan, Y. Cao, J. Liang, Y. Cao, M.Y. Yang, Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection, Inf. Fusion 50 (2019) 148–157, http://dx.doi.org/10.1016/j.inffus.2018.11.017.

[18] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26, http://dx.doi.org/10.1016/j.inffus.2018.09.004.

[19] L. Tang, J. Yuan, J. Ma, Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network, Inf. Fusion 82 (2022) 28–42, http://dx.doi.org/10.1016/j.inffus.2021.12.004.

[20] L. Tang, J. Yuan, H. Zhang, X. Jiang, J. Ma, PIAFusion: A progressive infrared and visible image fusion network based on illumination aware, Inf. Fusion 83 (2022) 79–92, http://dx.doi.org/10.1016/j.inffus.2022.03.007.

[21] L. Tang, H. Zhang, H. Xu, J. Ma, Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity, Inf. Fusion 99 (2023) 101870, http://dx.doi.org/10.1016/j.inffus.2023.101870.

[22] A. Mosella-Montoro, J. Ruiz-Hidalgo, 2D–3D geometric fusion network using multi-neighbourhood graph convolution for RGB-d indoor scene classification, Inf. Fusion 76 (2021) 46–54, http://dx.doi.org/10.1016/j.inffus.2021.05.002.

[23] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images., ECCV (5) 7576 (2012) 746–760.

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

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

[26] G. Franchi, M. Hariat, X. Yu, N. Belkhir, A. Manzanera, D. Filliat, InfraParis: A multi-modal and multi-task autonomous driving dataset, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, IEEE, 2024, pp. 2973–2983, http://dx.doi.org/10.1109/WACV56688.2024.00304.

[27] M. Baltaxe, T. Pe'er, D. Levi, Polarimetric imaging for perception, in: Proceedings of the 34th British Machine Vision Conference, BMVC, British Machine Vision Association, 2023, p. 566, URL https://proceedings.bmvc2023.org/566/.

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

[29] H. Teng, Y. Wang, X. Song, K. Karydis, Multimodal dataset for localization, mapping and crop monitoring in citrus tree farms, in: G. Bebis, G. Ghiasi, Y. Fang, A. Sharf, Y. Dong, C. Weaver, Z. Leo, J.J. LaViola Jr., L. Kohli (Eds.), Advances in Visual Computing, Springer Nature Switzerland, Cham, 2023, pp. 571–582, http://dx.doi.org/10.1007/978-3-031-47969-4_44.

[30] M.F. Kragh, P. Christiansen, M.S. Laursen, M. Larsen, K.A. Steen, O. Green, H. Karstoft, R.N. Jørgensen, Fieldsafe: Dataset for obstacle detection in agriculture, Sensors 17 (11) (2017) http://dx.doi.org/10.3390/s17112579, URL http://www.mdpi.com/1424-8220/17/11/2579.

[31] M. Abdulsalam, Z. Chekakta, N. Aouf, M. Hogan, Fruity: a multi-modal dataset for fruit recognition and 6D-pose estimation in precision agriculture, in: 2023 31st Mediterranean Conference on Control and Automation, MED, IEEE, 2023, pp. 144–149, URL https://github.com/MahmoudYidi/Fruity.

[32] P.J. Navarro, L. Miller, M.V. Díaz-Galián, A. Gila-Navarro, D.J. Aguila, M. Egea-Cortines, A novel ground truth multispectral image dataset with weight, anthocyanins, and brix index measures of grape berries tested for its utility in machine learning pipelines, GigaScience 11 (2022) giac052.

[33] 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. Mechatronics (2022).

[34] J. Strohmayer, M. Kampel, A compact tri-modal camera unit for RGBDT vision, in: 2022 the 5th International Conference on Machine Vision and Applications, ICMVA, in: ICMVA 2022, 2022, pp. 34–42, http://dx.doi.org/10.1145/3523111.3523116.

[35] Y. Choi, N. Kim, S. Hwang, K. Park, J.S. Yoon, K. An, I.S. Kweon, KAIST multi-spectral day/night data set for autonomous and assisted driving, IEEE Trans. Intell. Transp. Syst. 19 (3) (2018) 934–948.

[36] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, S. Marcel, Biometric face presentation attack detection with multi-channel convolutional neural network, IEEE Trans. Inf. Forensics Secur. 15 (2019) 42–55.

[37] A. George, S. Marcel, Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks, IEEE Trans. Inf. Forensics Secur. 16 (2020) 361–375.

[38] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph. Image Process. 39 (3) (1987) 355–368, http://dx.doi.org/10.1016/S0734-189X(87)80186-X.

[39] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: P.S. Heckbert (Ed.), Graphics Gems IV, Academic Press, San Diego, CA, 1994, pp. 474–485.

[40] H.I. Ashiba, H.M. Mansour, H.M. Ahmed, M.I. Dessouky, M.F. El-Kordy, O. Zahran, F.E.A. El-Samie, Enhancement of infrared images using histogram processing and the undecimated additive wavelet transform, Multimedia Tools Appl. 78 (9) (2019) 11277–11290, http://dx.doi.org/10.1007/s11042-018-6545-9.

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

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

[43] M.P. Das, L.H. Matthies, S. Daftry, Online photometric calibration of automatic gain thermal infrared cameras, IEEE Robot. Autom. Lett. 6 (2) (2021) 2453–2460, http://dx.doi.org/10.1109/LRA.2021.3060149.

[44] W. Jamrozik, J. Górka, G.F. Batalha, Dynamic range compression of thermograms for assessment of welded joint face quality, Sensors 23 (4) (2023) 1995, http://dx.doi.org/10.3390/s23041995.

[45] D.-G. Lee, J. Kim, Y. Cho, A. Kim, Thermal chameleon: Task-adaptive tone-mapping for radiometric thermal-infrared images, 2024, ArXiv:2410.18340 [Cs.CV].

[46] A. Gödrich, D. König, G. Eilertsen, M. Teutsch, Joint tone mapping and denoising of thermal infrared images via multi-scale retinex and multi-task learning, in: Infrared Technology and Applications XLIX (Proc. SPIE), 12534, 2023, 1253417, http://dx.doi.org/10.1117/12.2663745.

[47] H. Li, S. Wang, S. Li, H. Wang, S. Wen, F. Li, Thermal infrared image enhancement algorithm based on multi-scale guided filtering, Fire 7 (6) (2024) 192, http://dx.doi.org/10.3390/fire7060192.
[48] Y. Zhu, Y. Zhou, W. Jin, L. Zhang, G. Wu, Y. Shao, A low-delay dynamic range compression and contrast enhancement algorithm based on an uncooled infrared sensor with local optimal contrast, Sensors 23 (21) (2023) 8860, http://dx.doi.org/10.3390/s23218860.
[49] M.M. Duch, J.R. Morros, J. Ruiz-Hidalgo, Depth map compression via 3D region-based representation, Multimedia Tools Appl. 76 (11) (2017) 13761–13784.
[50] T. Sonoda, A. Grunnet-Jepsen, Depth image compression by colorization for intel RealSense depth cameras, 2021, Intel RealSense White Paper https://dev.intelrealsense.com/docs/depth-image-compression-by-colorization-for-intel-realsense-depth-cameras.
[51] J. Rambach, B. Mirbach, Y. Anisimov, D. Stricker, Time-of-flight depth sensing for automotive safety and smart building applications: The VIZTA project, IEEE Access 11 (2023) 105819–105829, http://dx.doi.org/10.1109/ACCESS.2023.3320268.
[52] M. Chen, P. Zhang, Z. Chen, Y. Zhang, X. Wang, S. Kwong, End-to-end depth map compression framework via RGB-to-depth structure priors learning, in: IEEE International Conference on Image Processing, ICIP, Bordeaux, France, 2022, pp. 3206–3210.
[53] J.-P. D'Amato, FitDepth: Fast and lite 16-bit depth image compression algorithm, EURASIP J. Image Video Process. 2023 (2023) 5.
[54] A.D. Wilson, Fast lossless depth image compression, in: Proc. ACM International Conference on Interactive Surfaces and Spaces, ISS, 2017, pp. 100–105.
[55] M.A. Tahouri, A.A. Alecu, L. Denis, A. Munteanu, Lossless and near-lossless L_∞ compression of depth video data, Sensors 25 (5) (2025) 1403.
[56] R. Gopalapillai, D. Gupta, M. Zakariah, Y.A. Alotaibi, Convolution-based encoding of depth images for transfer learning in RGB-d scene classification, Sensors 21 (23) (2021) 7950.
[57] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-D images for object detection and segmentation, in: European Conference on Computer Vision, ECCV, Springer, 2014, pp. 345–360.
[58] S.H. Kumar, K.R. Ramakrishnan, Depth compression via planar segmentation, Multimedia Tools Appl. 78 (5) (2019) 6529–6558.
[59] F. Tan, Z. Xia, Y. Ma, X. Feng, 3D sensor based pedestrian detection by integrating improved HHA encoding and two-branch feature fusion, Remote. Sens. 14 (3) (2022) 645.
[60] P. Ruiu, L. Mascia, E. Grosso, Saliency-guided point cloud compression for 3D live reconstruction, Multimodal Technol. Interact. 8 (5) (2024) 36.
[61] Z. Zhang, A flexible new technique for camera calibration, IEEE Trans. Pattern Anal. Mach. Intell. 22 (11) (2000) 1330–1334.
[62] G. Bradski, A. Kaehler, OpenCV library, 2000, (Accessed 22 March 2023), https://opencv.org/.
[63] L. Grady, Random walks for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1768–1783, http://dx.doi.org/10.1109/TPAMI.2006.233.
[64] J.F. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8 (6) (1986) 679–698, http://dx.doi.org/10.1109/TPAMI.1986.4767851.
[65] S. van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, the scikit-image contributors, Scikit-image: Image processing in python, 2014, URL https://scikit-image.org/ Version 0.10.0.
[66] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[67] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: European Conference on Computer Vision, ECCV, 2014, pp. 740–755.
[68] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2961–2969.
[69] M.J. Sousa, A. Moutinho, M. Almeida, Thermal infrared sensing for near real-time data-driven fire detection and monitoring systems, Sensors 20 (23) (2020) 6803.
[70] 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.
[71] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vis. Graph. Image Process. 39 (3) (1987) 355–368.
[72] V.E. Vickers, Plateau equalization algorithm for real-time display of high-quality infrared imagery, Opt. Eng., Bellingham 35 (7) (1996) 1921–1926, http://dx.doi.org/10.1117/1.601006.
[73] D.J. Jobson, Z.-u. Rahman, G.A. Woodell, A multiscale retinex for color image enhancement, IEEE Trans. Image Process. 6 (7) (1997) 965–976.
[74] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[75] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Stat. 27 (3) (1956) 832–837, http://dx.doi.org/10.1214/aoms/1177728190.
[76] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (3) (1962) 1065–1076, http://dx.doi.org/10.1214/aoms/1177704472.
[77] F. Tan, Z. Xia, Y. Ma, X. Feng, 3D sensor based pedestrian detection by integrating improved hha encoding and two-branch feature fusion, Remote. Sens. 14 (3) (2022) http://dx.doi.org/10.3390/rs14030645, URL https://www.mdpi.com/2072-4292/14/3/645.
[78] L.S. Team, Label Studio: Open Source Data Labelling Platform, 2024, URL https://labelstud.io/ (Accessed 8 February. 2025).
[79] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. Berg, W.-Y. Lo, P. Dollár, K. He, Segment anything, 2023, ArXiv Preprint arXiv:2304.02643 URL https://arxiv.org/abs/2304.02643.
[80] HumanSignal, Label studio SDK: Python client for label studio API, 2025, (Accessed 4 April 2025) https://github.com/HumanSignal/label-studio-sdk.