Pre-Logit Decoder Fusion for Five-Modality Segmentation with Unaligned T/UV Auxiliaries

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*

Abstract

We investigate decoder-level multimodal fusion for semantic segmentation with unaligned modalities (RGB+DIN(depth-intensity-normals), thermal, and ultraviolet (UV)). We introduce Cross-Modal Attention with Gated Residuals (CMAG), a hybrid module operating at the pre-logit stage that combines two complementary pathways: Global Context Modality Attention (GCMA), which establishes soft correspondences between thermal/UV and RGB+DIN features, and sigmoid-gated (SIG) residuals that inject per-pixel corrections from auxiliary modalities. Independent decoders generate quarter-resolution pre-logits per modality, preserving modularity while enabling robust handling of missing inputs, thereby eliminating explicit calibration requirements. We implement five decoder-level baselines, adapting established fusion paradigms to contextualise CMAG's performance.

On the MM5 dataset, CMAG achieves 84.18% mIoU across lighting conditions (underexposed/ideal/overexposed: 82.54%/87.61%/82.38%), outperforming attention-only GCMA (80.49%/86.72%/78.03%). Ablations reveal the importance of hierarchical modality: RGB and DIN removal cause severe degradation (59.50 pp and 49.61 pp, respectively), while thermal and UV provide specialised cues (24.62pp and 16.82pp losses). Spatial misalignment proves substantially less damaging than modality removal (20-pixel shifts: 2.61 pp vs 37.64 pp for drops), validating decoder fusion's alignment tolerance. Architectural comparison reveals distinct robustness profiles: CMAG maximises clean-data accuracy but shows elevated noise sensitivity (12.93 pp mean degradation), adapted Multimodal Transfer Module (PL-MMTM) achieves superior modality-drop robustness (31.82 pp), and adapted Recurrent Attention U-Net (PL-R2AU) demonstrates best noise resilience (8.80 pp). Comparison with encoder-level fusion (GF-Net) reveals fundamental trade-offs: encoder fusion achieves +1.34 pp (vs CMAG) accuracy and about 2× throughput, but suffers severe sensor-failure degradation, while decoder fusion prioritises robustness through late integration. These findings establish decoder-level fusion as viable for unaligned multimodal segmentation when robustness outweighs peak performance.

Keywords: Multimodal fusion, Thermal imaging, UV imaging, Late fusion, Sensor fusion, Semantic segmentation, Vision Transformers, Real-time fusion

1. Introduction

Robust semantic segmentation in field robotics, industrial inspection, and agriculture requires tolerance to fluctuating illumination, reflective or transparent materials, occlusions, and camouflage-like textures. Single-stream RGB systems are brittle under such conditions. In contrast, complementary sensors provide cues that compensate for specific weaknesses: depth supplies geometry independent of colour, thermal imaging highlights temperature-emissive regions, infrared intensity broadens the dynamic range, near-infrared (NIR) extends the observable spectrum, and ultraviolet (UV) reveals fluorescence and sub-visible surface structure [**?**]. Harnessing this heterogeneity promises finer delineation and more reliable decisions across domestic, industrial, and agricultural settings.

*Corresponding author
Email address: martin.brenner.1@uni.massey.ac.nz (Martin Brenner)

In practical capture rigs, not all streams are geometrically compatible. In our setup, RGB-D-NIR are inherently co-registered by the factory-calibrated RGB-D sensor, whereas UV and thermal (LWIR) use different optics, exhibit lens distortion, and are unaligned with respect to RGB-D-NIR. Early (data-level) and intermediate (feature-level) fusion strategies typically assume either prior geometric registration or rely on learnt, in-network rectification. Examples include CMX's Cross-Modal Feature Rectification Module (FRM) [**?**] and trimodal encoder attention in ETFormer [**?**], which adapt one stream using another before mixing. These mechanisms increase complexity and can remain brittle under residual misalignment or viewpoint change, which partly explains the prevalence of two-stream pairings (RGB-D, RGB-T) and the limited scalability to larger, heterogeneous sensor sets.

We propose CMAG (Cross-Modal Attention with Gated Residuals), a decoder-level fusion module that integrates unaligned modalities at the pre-logit stage through two complementary mechanisms. The primary path (RGB augmented with DIN (Depth, Intensity, and Normals [**?**]) forms the base representation, while LWIR thermal (T24) and UV (U8) are processed through separate encoder-decoder branches, preserving their native geometry. At the decoder's pre-logit stage, GCMA (Global Context Modality Attention) performs efficient cross-modal attention by querying thermal/UV features with RGB-DIN representations using downsampled tokens, extracting global context without requiring explicit spatial calibration. CMAG then applies sigmoid-gated (SIG) residuals, injecting per-pixel auxiliary modality corrections directly in feature space before final classification through a lightweight 1×1 convolution. Unlike encoder-level fusion, which presupposes spatial alignment, CMAG operates directly on unaligned sensor streams and remains compatible with optional alignment modules when available (Figure 1). Additionally, we introduce MWPA (Modality-Wise Parallel Attention). This computationally more efficient alternative employs parallel channel and spatial attention mechanisms for modality-selective fusion, providing a lightweight baseline for comparative analysis of decoder-level fusion strategies.

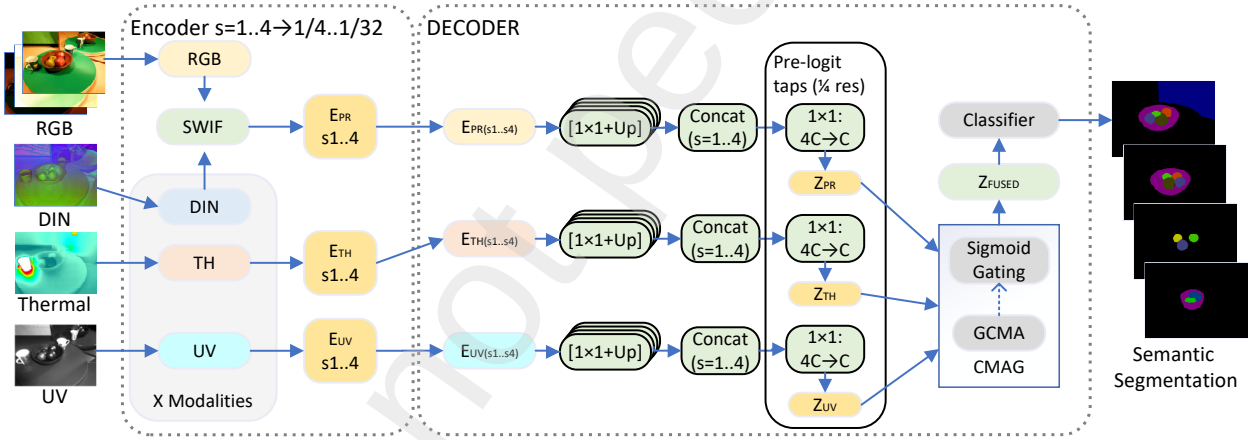


Figure 1: Overview of CMAG with three streams (RGB+DIN, LWIR, UV). At each encoder stage $s \in \{1, 2, 3, 4\}$, the primary path applies SWIF [**?**] to enhance RGB with DIN, yielding $E_{\text{PR}}^{(s)}$; thermal and UV produce $E_{\text{TH}}^{(s)}$ and $E_{\text{UV}}^{(s)}$ while preserving native geometry. Each decoder consumes its stage bundle $\{E^{(s)}\}$, applies [1×1 + Up→$\frac{1}{4}$] per stage, Concats over s=1..4, then reduces channels with 1×1 : $4C \to C$ (SE optional) to form the pre-logits $Z_{\text{PR}}, Z_{\text{TH}}, Z_{\text{UV}}$ at $\frac{1}{4}$ resolution. CMAG fuses these pre-logits: GCMA attends from Z_{PR} to $(Z_{\text{TH}}, Z_{\text{UV}})$ to produce F_{gcma}, and sigmoid-gated residuals add thermal/UV contributions to obtain $Z_{\text{fused}} = F_{\text{gcma}} + r_{\text{TH}} + r_{\text{UV}}$. A 1×1 classifier and ×4 upsampling yield fused logits at $H \times W$ without explicit geometric warping of LWIR/UV. Here, s indexes resolutions from $\frac{1}{4}$ to $\frac{1}{32}$; C denotes the decoder channel width.

To establish comprehensive decoder-level benchmarks and contextualise CMAG's performance, we adapt established fusion paradigms, PL-MMTM and PL-R2AU, to our alignment-free framework. PL-MMTM extends the channel-wise squeeze-and-excitation mechanism [**?**] to pre-logit features, enabling cross-modal salience transfer without spatial correspondence. PL-R2AU adapts recurrent attention gates [**?**] for consistent spatial focus across modalities. We evaluate GCMA and sigmoid gating as standalone modules to isolate the constituent mechanisms of CMAG. This unified framework, comprising six decoder-level variants, enables a systematic assessment of fusion complexity versus performance trade-offs, ranging from lightweight gating (PL-SIG) and parallel attention (MWPA) to channel modulation (PL-MMTM), recurrent spatial attention (PL-R2AU), and global cross-modal attention (GCMA, CMAG). To support training and evaluation in this alignment-free regime, we curate a new MM5 subset comprising raw,

unaligned RGB-D-NIR-T-UV imagery. Thermal and UV undergo only coarse preprocessing to establish overlapping fields of view, followed by random crops that mimic realistic misalignment and background variation (Section 3). This subset complements the aligned MM5 release [**?**].

1.1. Key Contributions

The main contributions of this work are:

1. We introduce CMAG, a decoder-level fusion module that integrates unaligned thermal and UV streams into an RGB-Depth-Intensity-Normals backbone via global cross-modal attention and sigmoid-gated residuals, achieving alignment-tolerant fusion without explicit geometric calibration.
2. We design a family of decoder-level fusion baselines by adapting established encoder/feature-level mechanisms (MMTM, R2AU, GF-Net style sigmoid gating and Modality-Wise Parallel Attention) to the pre-logit stage, enabling controlled, like-for-like comparisons of gating versus attention under a shared backbone and training protocol.
3. We conduct an extensive robustness characterisation, systematically evaluating all six decoder-level fusion architectures across modality dropout, spatial misalignment, and sensor noise injection, with a systematic comparison against encoder-level fusion to quantify the impact of fusion stage choice on accuracy-robustness trade-offs.
4. We curate and release an unaligned MM5 subset with raw, lens-distorted thermal and UV imagery, together with code and trained weights for CMAG and all decoder-level baselines, providing a reproducible benchmark for multimodal segmentation with misaligned auxiliary sensors.

2. Related Work

Decoder-Level Fusion Methods. Decoder-level fusion integrates modalities during the upsampling phase, before final classification. Three architectural patterns dominate this space. Hyper-fusion decoders aggregate multi-scale features from separate encoders: OctopusNet merges per-modality pyramids at each decode stage [**?**]. Multi-branch decoders maintain separate pathways: Mirror U-Net pairs modality-specific decoders with an auxiliary multimodal decoder and consistency objectives [**?**]. Central fusion decoders route all modalities through shared layers: SGFNet derives semantic guidance maps to reweight features [**?**], MEFNet employs modality experts [**?**], and GMFNet applies pixel-wise gates within U-shaped decoders [**?**].
Whilst these architectures demonstrate the effectiveness of feature-level integration within decoder structures, recent work has explored the extreme of purely decision-level approaches. LF-DLM reports approximately 0.3% mIoU gains by fusing modality-specific logits while keeping encoders and decoders completely independent [**?**]. This minimal improvement establishes a crucial baseline, highlighting that the marginal benefits of pure late fusion motivate exploration of more sophisticated cross-modal mechanisms that can capture richer inter-modality relationships without sacrificing the modularity advantages of decoder-level architectures.

Cross-Modal Attention for Fusion. Building upon the insights from decoder-level fusion limitations, cross-modal attention mechanisms have emerged as a powerful paradigm for integrating heterogeneous streams without requiring strict spatial alignment. CMAF-Net embeds cross-modal attention within a multi-encoder 3D U-Net, learning modality-invariant latents from incomplete multi-sequence MRI [**?**]. The CMAF block achieves strong Dice scores on BraTS 2020 under conditions of missing modality, demonstrating robustness to incomplete data. CMNeXt scales to arbitrary modalities on DeLiVER through late, lightweight attention-based integration with minimal per-modality parameters [**?**]. CANet employs bidirectional co-attention between RGB and depth features [**?**], while UCTNet incorporates uncertainty-aware cross-modal transformers [**?**]. These attention-based approaches offer dynamic, learnable mechanisms for modality interaction that adapt to input characteristics, addressing the limitations of static fusion rules observed in pure decision-level methods.

Attention-Based Feature Refinement. Whilst cross-modal attention facilitates inter-modality exchange, complementary research has focused on refining individual modality representations through channel and spatial recalibration before fusion occurs. CBAM applies sequential attention [**?**], while BAM employs parallel branches [**?**]; both operate on unimodal features to enhance their discriminative power. In multimodal contexts, this intra-modal refinement proves particularly valuable: MEFNet introduces modality-specific expert networks with attention-based reweighting [**?**], and TriFuse applies tri-attention across spatial, channel, and modality dimensions [**?**]. These refinement strategies

recognise that optimal fusion requires not only effective cross-modal interaction but also maximally informative individual modality representations. Unlike methods that process modalities independently or via cross-attention [**?**], our MWPA bridges these paradigms by deriving per-modality attention weights from concatenated representations, enabling simultaneous recalibration of both channels and spatial dimensions before fusion.

Gating Mechanisms. Whilst attention mechanisms recalibrate features, learnt gating directly assigns per-pixel or per-channel reliability weights to modulate modality contributions. MMTM implements channel-wise squeeze-and-excitation for cross-modal salience transfer [**?**]. SSMA demonstrates sigmoid-style recalibration for adaptive feature scaling [**?**]. GMFNet embeds pixel-wise gates in U-shaped decoders [**?**]. DGFNet's dual-gate design merges spatial detail with semantics during decoding [**?**]. R2AU-Net incorporates recurrent attention gates for consistent spatial focus [**?**].
Quality-aware gating extends basic reliability weighting. QSF-Net estimates quality maps for trimodal visible-depth-thermal integration [**?**]. MGFNet applies lightweight gating for optical-SAR fusion under cross-sensor noise [**?**]. These mechanisms prove valuable when modality quality varies or when inputs misalign.

Feature-Level Fusion Baselines. In contrast to decoder-level gating approaches, encoder-level fusion methods integrate modalities earlier in the network, achieving strong performance when inputs align. CMX's cross-modal feature rectification module (FRM) calibrates features before mixing [**?**]. ETFormer demonstrates single-encoder multimodal attention for RGB-D-T through task-specific pretraining [**?**]. GF-Net performs early RGB enhancement using SWIF and per-pixel sigmoid gating at encoder stages [**?**] for semantic segmentation using five modality fusion. These methods assume or restore spatial correspondence before decoding, limiting their applicability to unaligned sensors.

2.1. Alignment Handling in Decoder Fusion

The alignment dependency of encoder-level methods motivates decoder fusion approaches that accommodate misregistration through soft alignment via cross-modal attention or fusion at coarser scales, where parallax reduces [**? ?**]. When stronger consistency is required, feature rectification precedes late fusion (e.g., CMX's FRM [**?**]). Performance differences between fusion points depend on the dataset and task rather than a universal ranking [**? ?**].
Recent methods directly target misalignment during decoding. Project-and-Fuse uses texture-prior guidance to mitigate biassed assignment in unaligned streams without pixel-exact registration, optimising a divergence-based loss to discourage degenerate assignments [**?**]. LMFNet employs lightweight transformer decoders that attend to heterogeneous cues across scales, while keeping compute manageable for high-resolution inputs [**?**].

2.2. Uncertainty and Dynamic Reliability in Decoder Fusion

Beyond handling spatial misalignment, dynamic reliability modelling at the decoder extends basic gating mechanisms. UDFNet couples uncertainty estimation with dynamic fusion, reporting strong accuracies across Berlin, Augsburg, MUUFL and Trento benchmarks, outperforming prior decoder-only fusion by sizeable margins [**?**]. However, the strongest results in overhead remote sensing often come from hybrid pipelines blending feature-level and decision-level fusion (e.g., MCAM/CPS modules with decision fusion in TCPSNet, prototype compensation in PICNet), indicating room for improvement in decoder-only strategies [**?**].

2.3. Positioning of CMAG

CMAG addresses decoder-level fusion for unaligned, heterogeneous sensors. Existing decoder-level methods assume aligned inputs [**? ?**] or use coarse decision-level fusion [**?**]. Encoder-level approaches [**? ?**] typically operate on stereo-calibrated, pre-registered datasets and employ feature rectification to refine alignment at the encoder level. Explicit spatial warping via parametric transformations [**? ?**] can accommodate geometric distortions but introduces computational overhead. In contrast, CMAG operates directly on unaligned auxiliary modalities at the decoder level, leveraging cross-modal attention for alignment-tolerant fusion without explicit spatial transformation.
Architecturally, CMAG performs single-stage fusion at the pre-logit level after multi-scale feature aggregation. Global Context Modality Attention (GCMA) establishes soft correspondence by attending to pooled modality representations rather than dense spatial tokens, accommodating misalignment with reduced complexity. Sigmoid-gated residuals add fine-grained spatial corrections. This design contrasts with per-stage fusion [**?**], concatenation-based methods [**? ?**], and full spatial attention approaches [**?**].

Separate per-modality decoder heads provide granular supervision and enable more graceful degradation under missing inputs, while learnt gate maps offer spatial interpretability of modality contributions. This modularity distinguishes CMAG from single-head architectures and channel-wise methods [**?**].
Table 1 summarises the supervision strategies and architectural choices across these methods, highlighting the diversity of approaches to modality-specific training and inference.

Table 1: Overview of modality-specific heads and GT usage in representative multimodal fusion methods. "Mod.-specific heads" = separate output heads per modality during training; "Mod.-specific GT" = different targets per modality.

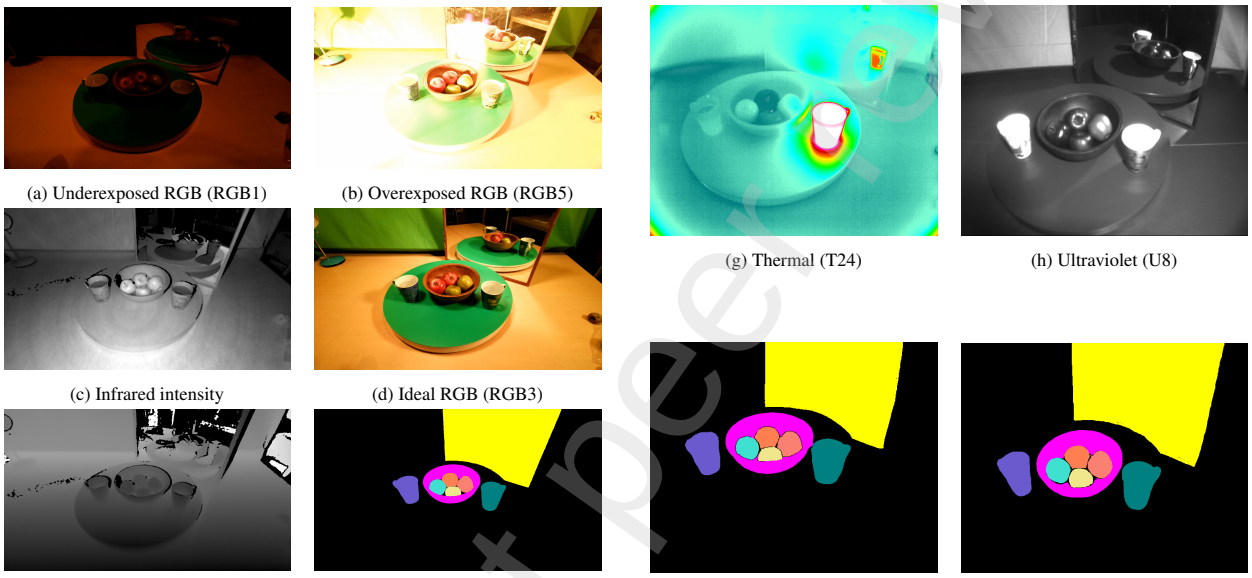
Method	**Year**	**Modalities**	**Fusion category**	**Mod.-specific heads**	**Mod.-specific GT**	**Supervision summary**	**Inference head(s)**
Mirror U-Net [**?**]	2023	PET+CT; MRI sequences	Decoder feature (multi-branch)	Yes	Yes	Modality-specific decoders with task-tailored supervision; auxiliary multimodal decoder.	Fused/combined
GMFNet [**?**]	2020	RGB+T	Decoder feature (central+lateral)	Yes	No	Two lateral unimodal + one central multimodal U-Net; shared semantic GT.	Central decoder
SGFNet [**?**]	2023	RGB+T	Decoder feature (central)	No	No	Semantic guidance maps reweight decoder features	single
MEFNet [**?**]	2023	RGB+Thermal	Decoder feature (experts)	No	No	Modality experts aggregated in shared decoder	single
DGFNet [**?**]	2021	RGB+T	Decoder feature (dual-gate)	No	No	Positional and filter gates merge spatial/channel attention	single
R2AU-Net [**?**]	2021	RGB+T	Decoder feature (recurrent attention)	No	No	Recurrent attention gates for consistent spatial focus	single
OctopusNet [**?**]	2019	Multi-contrast MRI	Decoder feature (hyper-fusion)	No	No	Separate encoders; hyper-fusion decoder with multi-scale integration.	single
CMAF-Net [**?**]	2024	MRI (T1, T1ce, T2, FLAIR)	Decoder feature (attention)	No	No	Cross-modal attention fusion in 3D U-Net; handles missing modalities.	single
CMNeXt [**?**]	2023	Arbitrary modalities	Decoder attention	No	No	Self-query hub selects informative tokens per modality	single
QSF-Net [**?**]	2024	RGB+D+T	Hybrid (multi-stage)	Yes (stage-1)	No	Stage-wise: saliency, quality maps, fused saliency + edge.	single
UDFNet [**?**]	2025	HSI/SAR/LiDAR	Decoder feature (uncertainty)	No	No	Uncertainty-aware dynamic fusion at decode time	single
Project-and-Fuse [**?**]	2025	RGB+D (unaligned)	Alignment-free decoder	No	No	Texture-prior guided fusion; handles unaligned inputs	single
LF-DLM [**?**]	2024	VHR Aerial+Sentinel-2	Decision-level	Yes (per branch)	No	Per-branch probabilities fused by weighted geometric mean.	Combined outputs
CMAG (ours)	2025	RGB-Depth-NIR + Thermal + UV	Decoder feature (central fusion; single-pass pre-logit attention + gated residuals)	Yes	Yes	Separate heads for primary, thermal and UV plus a fused head; operates on unaligned, lens-distorted UV/LWIR without explicit calibration.	single (auxiliary feature paths active)

Note: For CMAG, thermal and UV decoder branches are executed at inference to produce pre-logit features consumed by the fusion module; their per-modality logits are not used for the final prediction (unless reported for diagnostics).

3. MM5 Dataset

The MM5 dataset [**?**], introduced in our prior work, was developed to address persistent limitations in existing multimodal segmentation benchmarks, most notably their constrained modality diversity, lack of raw sensor fidelity, and absence of unaligned annotation protocols. MM5 integrates five distinct imaging modalities: RGB, depth (D), thermal (T), ultraviolet (UV), and near-infrared (NIR) within a unified acquisition and annotation pipeline. The dataset was constructed using a custom acquisition rig built from off-the-shelf RGB-D sensors and complemented by thermographic and UV imaging systems. Unlike existing datasets that prioritise pre-registered inputs, MM5 provides both geometrically aligned and unaligned variants of each scene, along with pixel-level annotations. This dual-format

design supports research on both feature-level fusion and late and decoder-level fusion strategies, without requiring spatial registration.
Each scene in the MM5 dataset contains a varied selection of objects, encompassing fresh produce, plastic replicas, and partially decayed items. The scenes are captured under diverse illumination conditions, including shadows, underexposure, and saturation, ensuring that each modality offers distinct and complementary semantic information. In this work, MM5 serves as a critical resource, facilitating our exploration of decoder-level fusion strategies, particularly in scenarios characterised by imperfect or absent spatial alignment. Moreover, MM5 supports modality-specific supervision through the provision of independent ground truth annotations.
For our experiments on decoder-level fusion, we utilise the unaligned UV and thermal images, which inherently exhibit lens distortion, as illustrated in Figure 2. Conversely, depth and NIR images are spatially registered with RGB through the RGB-D sensor.



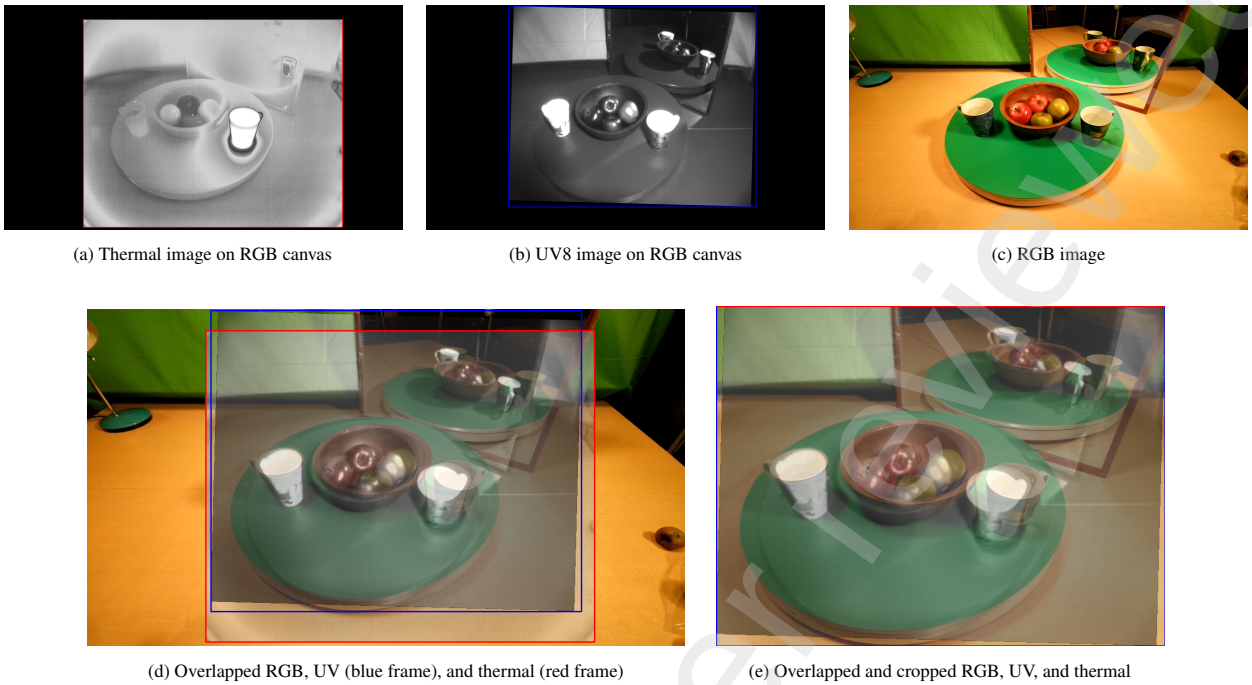
(a) Underexposed RGB (RGB1) (b) Overexposed RGB (RGB5) (g) Thermal (T24) (h) Ultraviolet (U8)
(c) Infrared intensity (d) Ideal RGB (RGB3)
(e) Depth (f) RGB ground truth mask (i) Thermal ground truth mask (j) UV ground truth mask

Figure 2: MM5 sample unaligned image subset for frame 544

Following the methodology established in our previous work, GF-Net [**?**], our analysis focuses specifically on underexposed, well-exposed, and overexposed RGB conditions to enable a direct and consistent comparison.
Due to variations in the image formats across modalities, we identified an approximate region of interest and uniformly cropped all images to a common maximum overlay area measuring 800 pixels in width and 600 pixels in height, as depicted in Figure 3. Subsequently, all labelled images were processed and assigned sequential numbering starting from 1, analogous to the aligned MM5 dataset, albeit with slight adjustments to folder naming conventions to avoid ambiguity. The class IDs and image IDs remain consistent with the aligned dataset.
Based on the *MM5_RAW_CROPPED* dataset, we created an additional dataset variant with reduced dimensions of 640 pixels in width and 480 pixels in height. To enhance background diversity and facilitate experiments involving camera misalignment, the cropping window position was varied randomly within the original frames. This approach enabled the generation of modality-specific pixel shifts, simulating realistic scenarios of camera displacement. Both datasets will be publicly released alongside the original MM5 dataset as subfolders *MM5_RAW_CROPPED* and *MM5_RAW_CRP640*, the latter includes a metadata file detailing the crop coordinates for each image.

3.1. Training and Evaluation Data

For comparability with previous work, we adopted the standard class-wise train-evaluation split provided by the MM5 dataset [**?**], as detailed in GF-Net [**?**]. This split maintains an approximate stratification, allocating around 75-80% of the class images to training, with the remaining 20-25% reserved for evaluation. Since the dataset contains



(a) Thermal image on RGB canvas

(b) UV8 image on RGB canvas

(c) RGB image

(d) Overlapped RGB, UV (blue frame), and thermal (red frame)

(e) Overlapped and cropped RGB, UV, and thermal

Figure 3: Overlap and cropping process for frame 544 of the MM5 raw data.

mainly mixed scenes, the distribution per class varies. However, the dataset exhibits significant class imbalance, with dominant categories such as Lemon and Mandarin having over a hundred annotated object instances, while others, such as Mandarin Peel and Kettle, are limited to a dozen or fewer examples. Certain classes, like Mandarin Peel, are particularly underrepresented, with evaluation sets as small as three images, while composite scenes introduce additional imbalances. To enhance training effectiveness and mitigate this imbalance, we identified scenes containing underrepresented classes and increased their frequency during training. This was achieved by applying diverse data augmentations, such as zoom, rotation, and flipping. Additionally, the dataset's long-tail class distribution provides an opportunity to evaluate and improve model robustness and generalisation across both rare and frequent classes alike. Further details on the exact composition, challenges, and rationale behind the dataset splits are provided in the original MM5 [**?**] and GF-Net [**?**] papers.

4. Proposed Methods

We adopt a decoder-centric fusion framework that performs efficient single-stage fusion at the pre-logit level. CMAG integrates two complementary mechanisms: Global Context Modality Attention (GCMA) for cross-modal feature exchange and per-pixel sigmoid gating for fine-grained spatial refinement. This design operates directly on unaligned sensor streams, establishing soft correspondence through learnt attention rather than explicit geometric alignment.

4.1. Architecture Overview

The network comprises three coordinated components:

1. **Primary RGB+DIN Stream (PR):** A MiT-B0 encoder enhanced at each stage by Stage-Wise Intensity Fusion (SWIF) [**?**], which injects pre-computed DIN composites (depth, infrared intensity, surface normals) [**?**] via lightweight residual blocks for lighting resilience.
2. **Auxiliary Thermal Stream (TH):** An independent MiT-B0 encoder processes long-wave infrared (LWIR) thermal imagery, preserving native geometry without spatial rectification.

3. **Auxiliary Ultraviolet Stream (UV):** An independent MiT-B0 encoder processes ultraviolet imagery, maintaining native geometry without spatial alignment.

4. **Pre-logit CMAG Fusion:** SegFormer-style MLP decoders applied to each stream produce 1/4-resolution pre-logit features ($Z_{\text{PR}}, Z_{\text{TH}}, Z_{\text{UV}}$). CMAG fuses these via global context attention and sigmoid-gated residuals to generate the final prediction.

Throughout this paper, we use the notation PR (Primary), TH (Thermal), and UV (Ultraviolet) to denote these three modality streams in equations and diagrams.

4.2. Decoder and Pre-logit Assembly

Multi-scale integration occurs within each stream's decoder. Each SegFormer-style MLP head projects its four encoder stages to a common embedding dimension C, upsamples them to a unified 1/4 scale, and aggregates via element-wise summation after 1×1 projection. This produces a single pre-logit feature tensor per stream:

$$Z_{\text{PR}},\ Z_{\text{TH}},\ Z_{\text{UV}} \ \in\ \mathbb{R}^{B\times C\times \frac{H}{4}\times \frac{W}{4}}, \tag{1}$$

where C=512 channels encode hierarchical features. Figure 4 illustrates this process.

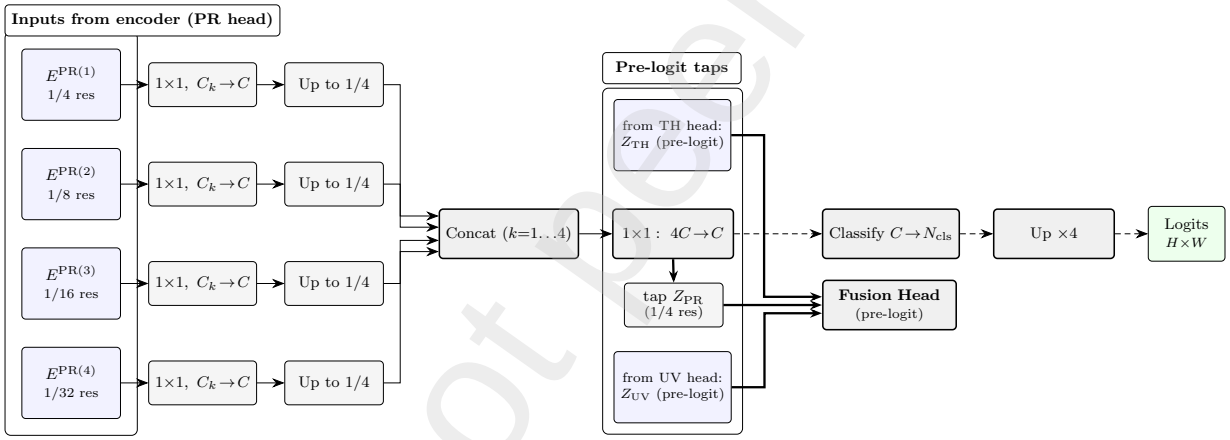


Figure 4: SegFormer-style MLP decoder for primary stream. Multi-scale encoder features ($E^{(0)}$ through $E^{(3)}$) are projected to C channels, upsampled to $H/4 \times W/4$, and summed to produce pre-logit features Z_{PR}. *Notation:* res = spatial resolution; C_k = stage width; C = decoder width; N_{cls} = number of classes. The dotted line indicates the single-head classification path.

4.3. Pre-logit CMAG Fusion

CMAG fuses the three pre-logit tensors through two sequential stages: global context modality attention (GCMA) followed by sigmoid-gated residuals (SIG).

4.3.1. Global Context Modality Attention (GCMA)

GCMA establishes cross-modal correspondence by operating at the modality level rather than the spatial-token level. Each pre-logit tensor is first lightly enhanced (Conv+Norm) and globally pooled to produce a modality context vector:

$$c_i \ =\ \text{GAP}(\text{Enhance}(Z_i)) \ \in\ \mathbb{R}^{B\times C}, \quad i \in \{\text{PR}, \text{TH}, \text{UV}\}. \tag{2}$$

Global Average Pooling (GAP). For $Z \in \mathbb{R}^{B\times C\times H\times W}$,

$$\text{GAP}(Z)_{b,c} \ =\ \frac{1}{HW}\sum_{y=1}^{H}\sum_{x=1}^{W} Z_{b,c,y,x} \ \in\ \mathbb{R}^{B\times C}, \tag{3}$$

i.e., one vector per sample and channel (no learnable parameters), aggregating spatial evidence into a single modality token per stream. The three modality contexts are stacked and processed by multi-head attention, with the primary context as the query:

$$
\begin{aligned}
C_{\text{stack}} &= \text{Stack}([c_{\text{PR}}, c_{\text{TH}}, c_{\text{UV}}]) \in \mathbb{R}^{B\times 3\times C},\\
Q &= \text{reshape}_h(W_Q\, c_{\text{PR}}) \in \mathbb{R}^{B\times h\times 1\times d_h},\\
K &= \text{reshape}_h(W_K\, C_{\text{stack}}) \in \mathbb{R}^{B\times h\times 3\times d_h},\\
V &= \text{reshape}_h(W_V\, C_{\text{stack}}) \in \mathbb{R}^{B\times h\times 3\times d_h},\\
&\quad \text{(with } h \text{ heads and per-head width } d_h\text{; typically } C = h\, d_h\text{)},\\
A &= \text{softmax}\Big(\tfrac{QK^\top}{\sqrt{d_h}}\Big) \in \mathbb{R}^{B\times h\times 1\times 3},\\
&\quad \text{(softmax over modalities)},\\
U &= AV \in \mathbb{R}^{B\times h\times 1\times d_h},\\
&\quad \text{(batched per-head matmul over the modality dimension)},\\
\hat{c} &= W_O\, \text{merge}_h(U) \;+\; c_{\text{PR}} \;\in\; \mathbb{R}^{B\times C},
\end{aligned}
\tag{4}
$$

so attention operates over M=3 modality tokens (not HW spatial tokens), giving $O(M^2)$ complexity versus $O(N^2)$ with $N{=}\frac{H}{4}\frac{W}{4}$ spatial tokens. The attended context is normalised and broadcast to pre-logit resolution:

$$
F_{\text{gcma}} \;=\; \text{Broadcast}(\text{Norm}(\hat{c})) \;\in\; \mathbb{R}^{B\times C\times \frac{H}{4}\times \frac{W}{4}}. \tag{5}
$$

Figure 5 illustrates this mechanism.

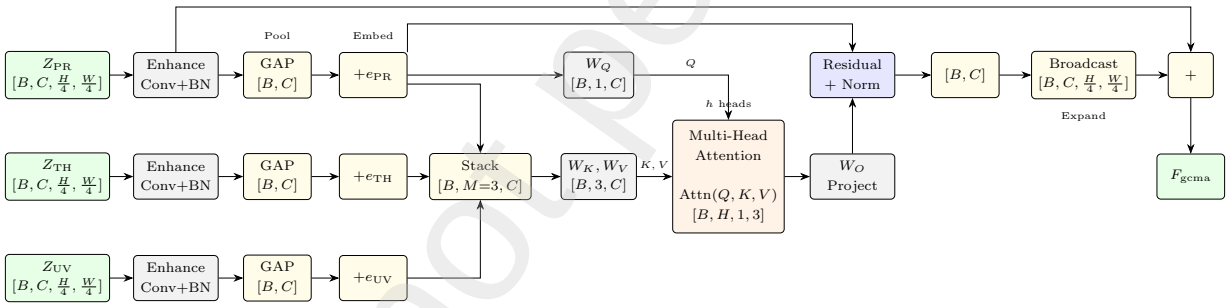


Figure 5: Global Context Modality Attention (GCMA). Pre-logits are enhanced and globally pooled to modality contexts $c_i \in \mathbb{R}^{B\times C}$, stacked to $\mathbb{R}^{B\times 3\times C}$. Multi-head attention computes $A = \text{softmax}(QK^\top/\sqrt{d_h})$ where query $Q \in \mathbb{R}^{B\times h\times 1\times d_h}$ from the primary stream attends to keys and values $K, V \in \mathbb{R}^{B\times h\times 3\times d_h}$ from all three modalities, producing attention weights $A \in \mathbb{R}^{B\times h\times 1\times 3}$ over the modality axis. The operation $U = AV$ denotes per-head batched matrix multiplication, computing an attention-weighted sum of V over the modality dimension. Output projection W_O merges the h heads (either $C{=}h\,d_h$ or $W_O : hd_h \to C$), residual connection with c_{PR}, and spatial broadcast yield $F_{\text{gcma}} \in \mathbb{R}^{B\times C\times H/4\times W/4}$.

4.3.2. Sigmoid-Gated Residuals (SIG)

Fine-grained spatial corrections are added via per-pixel sigmoid gates [**?**] that modulate transformed auxiliary features. For each auxiliary modality, a gate is computed from the concatenation of the GCMA context and the modality's pre-logit:

$$
\begin{aligned}
g_{\text{th}} &= \sigma(W_{\text{th}}^{(g)} * [\, F_{\text{gcma}} \,\|\, Z_{\text{TH}} \,]) \in \mathbb{R}^{B\times 1\times \frac{H}{4}\times \frac{W}{4}},\\
g_{\text{uv}} &= \sigma(W_{\text{uv}}^{(g)} * [\, F_{\text{gcma}} \,\|\, Z_{\text{UV}} \,]) \in \mathbb{R}^{B\times 1\times \frac{H}{4}\times \frac{W}{4}},\\
r_{\text{th}} &= g_{\text{th}} \odot (W_{\text{th}}^{(t)} * Z_{\text{TH}}), \quad r_{\text{uv}} = g_{\text{uv}} \odot (W_{\text{uv}}^{(t)} * Z_{\text{UV}}),\\
Z_{\text{fused}} &= F_{\text{gcma}} \;+\; r_{\text{th}} \;+\; r_{\text{uv}}.
\end{aligned}
\tag{6}
$$

The gate networks employ r=4 reduction for efficiency. A 1×1 classifier produces logits from Z_{fused}, which are upsampled to full resolution. Figure 6 shows the complete CMAG pipeline.

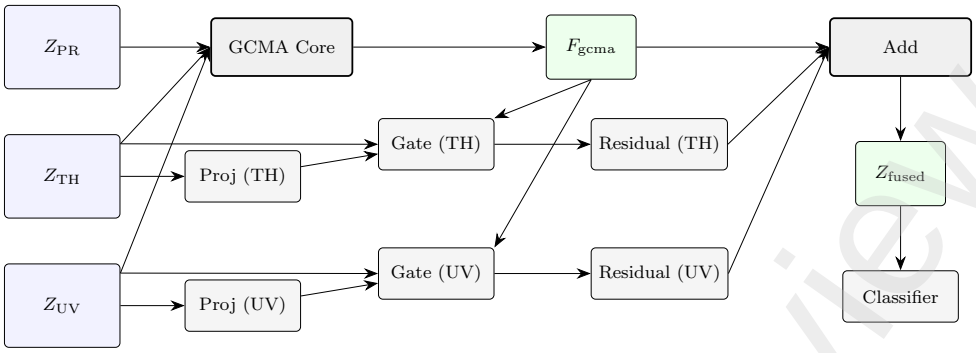


Figure 6: CMAG fusion overview. Multi-stage feature pyramids from three modalities are aggregated to pre-logit features ($Z_i \in \mathbb{R}^{B\times C\times H/4\times W/4}$) and projected to C=512 channels. GCMA pools each modality globally, applies multi-head attention over modality contexts (primary as query), and broadcasts to produce F_{gcma}. Sigmoid-gated residuals r_{TH} and r_{UV} are computed using spatial gates $g_{\mathrm{TH/UV}} \in [0,1]^{B\times 1\times H/4\times W/4}$. Final fusion: $Z_{\mathrm{fused}} = F_{\mathrm{gcma}} + r_{\mathrm{TH}} + r_{\mathrm{UV}}$.

4.4. Multi-Head Supervision

We train a fused head alongside three unimodal heads (primary, thermal, UV), all predicting the same semantic classes. Unimodal heads are supervised against modality-specific ground truth, while the fused head uses primary labels. The total objective:

$$\mathcal{L}_{\mathrm{total}} = \alpha_{\mathrm{pr}}\,\mathcal{L}_{\mathrm{pr}} + \alpha_{\mathrm{th}}\,\mathcal{L}_{\mathrm{th}} + \alpha_{\mathrm{uv}}\,\mathcal{L}_{\mathrm{uv}} + \alpha_{\mathrm{fused}}\,\mathcal{L}_{\mathrm{fused}}, \tag{7}$$

with non-negative weights α. This multi-head supervision provides fine-grained gradients to each stream while training the fusion mechanism via $\mathcal{L}_{\mathrm{fused}}$. At inference, auxiliary decoders produce pre-logits for fusion. By default, we output only the fused prediction, with optional per-head outputs for diagnostics at the cost of throughput.

4.5. MWPA (Modality-Wise Parallel Attention)

To isolate the contributions of cross-modal attention, we implement MWPA (Modality-Wise Parallel Attention), which applies parallel channel and spatial attention mechanisms rather than cross-modal feature exchange. This baseline enables systematic comparison of attention strategies while maintaining moderate computational efficiency. Given three pre-logit maps

$$Z_{\mathrm{PR}}, Z_{\mathrm{TH}}, Z_{\mathrm{UV}} \in \mathbb{R}^{B\times C\times H/4\times W/4}, \tag{8}$$

the method concatenates them along the channel dimension to form $X \in \mathbb{R}^{B\times 3C\times H/4\times W/4}$.

Channel Attention. The concatenated features undergo global average pooling followed by a two-layer MLP with reduction ratio r=16:

$$\mathbf{w}_c = \sigma(\mathrm{Conv}_{1\times 1}(\delta(\mathrm{Conv}_{1\times 1}(\mathrm{GAP}(X))))), \tag{9}$$

where the first convolution reduces from $3C$ to C/r channels, and the second expands back to $3C$ channels. The output $\mathbf{w}_c \in [0,1]^{B\times 3C\times 1\times 1}$ is reshaped to $[B, 3, C, 1, 1]$ to provide per-modality channel attention weights.

Spatial Attention. In parallel, the concatenated features are processed through a spatial attention network:

$$\mathbf{m}_s = \sigma(\mathrm{Conv}_{1\times 1}(\delta(\mathrm{Conv}_{3\times 3}(X)))), \tag{10}$$

where the 3×3 convolution (with padding=1) reduces from $3C$ to C/r channels, followed by a 1×1 convolution that produces N=3 spatial attention maps, yielding $\mathbf{m}_s \in [0,1]^{B\times 3\times H/4\times W/4}$.

Modality-Specific Weighting. Unlike sequential attention, this method applies both attention types simultaneously to each modality:

$$Z_m^{\mathrm{weighted}} = Z_m \odot \mathbf{w}_c^{(m)} \odot \mathbf{m}_s^{(m)}, \quad m \in \{\mathrm{PR}, \mathrm{TH}, \mathrm{UV}\}, \tag{11}$$

where $\mathbf{w}_c^{(m)} \in [0,1]^{B\times C\times 1\times 1}$ and $\mathbf{m}_s^{(m)} \in [0,1]^{B\times 1\times H/4\times W/4}$ are the channel and spatial attention weights for modality m.

Final Fusion. The weighted modalities are summed to produce the fused output:

$$Z_{\text{fused}} = \sum_{m\in\{\text{PR,TH,UV}\}} Z_m^{\text{weighted}} \in \mathbb{R}^{B\times C\times H/4\times W/4}. \tag{12}$$

This approach enables modality-specific attention learning while maintaining computational efficiency (Figure 7).

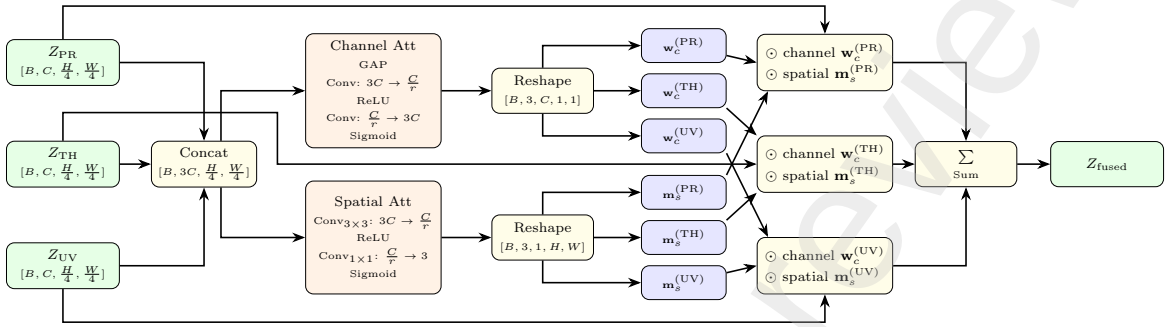


Figure 7: **Modality-wise Parallel Attention (MWPA).** From the concatenated pre-logits $[Z_{\text{PR}}\|Z_{\text{TH}}\|Z_{\text{UV}}]$, the module computes for each modality $m \in \{\text{PR, TH, UV}\}$:(i) a channel weight $\mathbf{w}_c^{(m)} \in [0,1]^{B\times C\times1\times1}$ via GAP→MLP (r=16),and (ii) a spatial mask $\mathbf{m}_s^{(m)} \in [0,1]^{B\times1\times\frac{H}{4}\times\frac{W}{4}}$ via 3×3/1×1 convolutions with sigmoid. Each modality is recalibrated as $\tilde{Z}_m = (\mathbf{w}_c^{(m)} \odot \mathbf{m}_s^{(m)}) \odot Z_m$ (element-wise with broadcasting), then summed to yield Z_{fused}. *Notation:* $\odot$ denotes element-wise (Hadamard) multiplication with broadcasting; applying channel then spatial (or vice versa) is equivalent.

5. Experimental Setup

5.1. Comparison Methods and Unified Evaluation Framework

To comprehensively evaluate CMAG, we establish a unified decoder-level fusion framework and adapt five representative methods spanning major fusion paradigms: (i) global channel modulation (MMTM [**?**]), (ii) recurrent spatial attention (R2AU [**?**]), (iii) cross-modal attention (GCMA), (iv) sigmoid gating (PL-SIG), and (v) hybrid attention (MWPA). These adaptations enable systematic comparison across diverse fusion paradigms within a consistent architectural framework.

Unified Framework. All methods operate within a standardised architecture: MiT-B0 backbone with three independent decoder heads (primary RGB+DIN, thermal, UV) producing C-channel pre-logit features at 1/4 spatial resolution. Fusion modules integrate these pre-logit features while maintaining consistent spatial dimensions and channel depth across all modalities. A single fusion operation occurs at the pre-logit stage, ensuring a fair comparison across methods. Table 3 reports parameter counts and computational complexity (FLOPs) to quantify the overhead introduced by different fusion mechanisms.

Adapted Implementations. Methods adapted to our decoder-level fusion framework are designated with a PL- prefix (Pre-Logit) to maintain a clear distinction from their original architectures. Each adaptation faithfully preserves the fundamental fusion mechanism while conforming to our standardised three-modality pre-logit integration scheme, enabling direct performance comparison across heterogeneous fusion strategies.

PL-MMTM adapts the multimodal transfer module [**?**] from its original two-stream architecture to our trimodal decoder framework via channel-wise squeeze-and-excitation. Given pre-logit features from the three decoder heads,

$$Z_{\text{PR}}, Z_{\text{TH}}, Z_{\text{UV}} \in \mathbb{R}^{B\times C\times H/4\times W/4}, \tag{13}$$

the method concatenates them along the channel dimension to form $Z_{\text{concat}} \in \mathbb{R}^{B\times3C\times H/4\times W/4}$. Global average pooling compresses spatial information into a channel descriptor $\mathbf{s} \in \mathbb{R}^{B\times3C}$. This descriptor undergoes bottleneck processing through sequential fully-connected layers with reduction ratio r=8:

$$\boldsymbol{\alpha} = \sigma(\mathbf{FC}_2(\delta(\mathbf{FC}_1(\mathbf{s})))), \quad \mathbf{FC}_1 : 3C \to \frac{3C}{r}, \quad \mathbf{FC}_2 : \frac{3C}{r} \to 3C, \tag{14}$$

where δ denotes ReLU and σ denotes sigmoid activation. The excitation weights $\boldsymbol{\alpha} \in [0,1]^{B\times 3C}$ are reshaped to $[B, 3C, 1, 1]$ and applied via element-wise multiplication to the concatenated features. The gated features are then split back into individual modalities $Z_{\text{PR}}^{\text{gated}}, Z_{\text{TH}}^{\text{gated}}, Z_{\text{UV}}^{\text{gated}} \in \mathbb{R}^{B\times C\times H/4\times W/4}$ and averaged to produce the final fused representation:

$$Z_{\text{fused}} = \frac{1}{3}(Z_{\text{PR}}^{\text{gated}} + Z_{\text{TH}}^{\text{gated}} + Z_{\text{UV}}^{\text{gated}}) \in \mathbb{R}^{B\times C\times H/4\times W/4}. \tag{15}$$

This approach enables each modality to be recalibrated based on the global context of all three inputs before fusion (Figure 8).

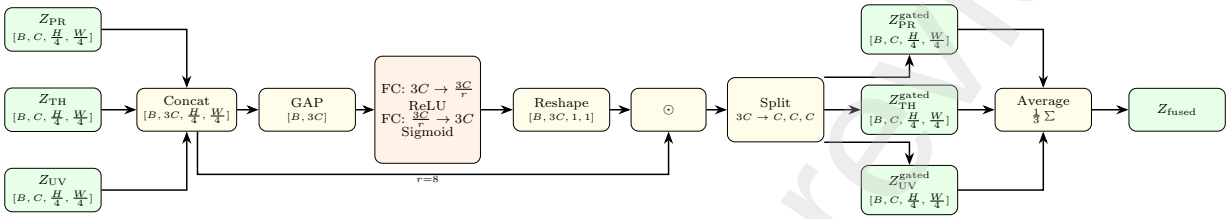


Figure 8: **PL-MMTM.** Trimodal fusion through channel-wise squeeze-and-excitation. Features are concatenated, globally pooled, and processed through an MLP bottleneck to generate channel excitation weights. After gating, features are split back to individual modalities and averaged to produce $Z_{\text{fused}} \in \mathbb{R}^{B\times C\times H/4\times W/4}$. *Notation:* $\odot$ denotes element-wise (Hadamard) multiplication with broadcasting.

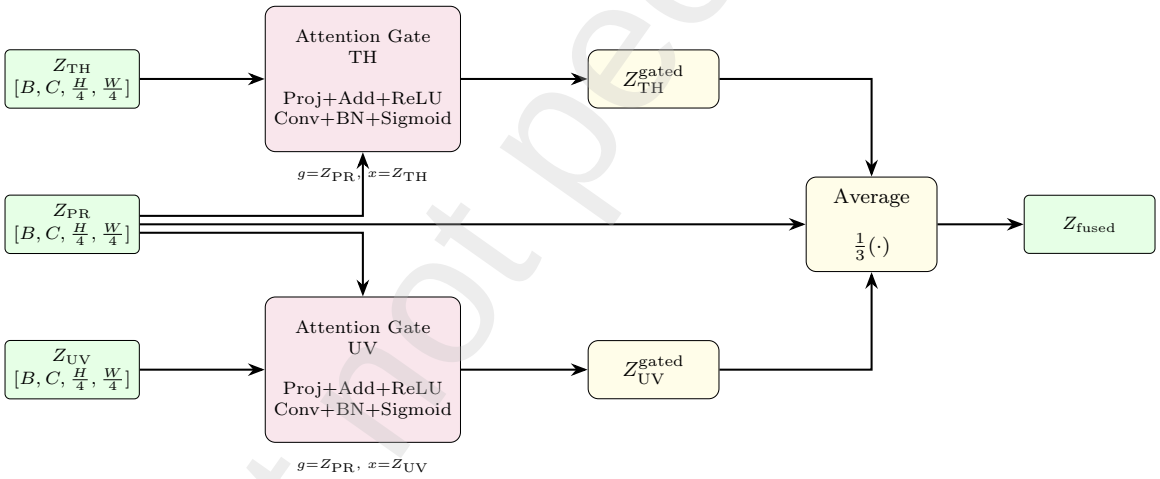


Figure 9: **PL-R2AU.** Primary features Z_{PR} serve as gating signals for auxiliary modalities Z_{TH} and Z_{UV} through attention gates. Each gate projects inputs to $F_{\text{int}} = C/4$, computes spatial attention masks via ReLU and sigmoid, then applies element-wise multiplication. The gated auxiliaries are averaged with the primary to yield Z_{fused}.

PL-R2AU adapts recurrent attention gates [**?**] to decoder-level fusion at the pre-logit stage. Given pre-logit features

$$Z_{\text{PR}}, Z_{\text{TH}}, Z_{\text{UV}} \in \mathbb{R}^{B\times C\times H/4\times W/4}, \tag{16}$$

the method employs two attention gates, each using the primary features Z_{PR} as the gating signal g to modulate an auxiliary modality x_m where $m \in \{\text{TH}, \text{UV}\}$. For each gate, both g and x_m undergo separate projections to an intermediate dimension $F_{\text{int}} = C/4$ via 1×1 convolutions with batch normalisation:

$$W_g : \mathbb{R}^C \to \mathbb{R}^{F_{\text{int}}}, \quad W_x : \mathbb{R}^C \to \mathbb{R}^{F_{\text{int}}}. \tag{17}$$

The projected features are element-wise summed and processed through ReLU activation, followed by a 1×1 convolution with batch normalisation and sigmoid activation to generate spatial attention masks:

$$\psi_m = \sigma(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathrm{ReLU}(W_g(g) + W_x(x_m))))) \in [0,1]^{B\times1\times H/4\times W/4}. \tag{18}$$

Each auxiliary modality is gated by its corresponding attention mask: $Z_m^{\text{gated}} = \psi_m \odot Z_m$. The final fused representation averages the primary features with the two gated auxiliaries:

$$Z_{\text{fused}} = \frac{1}{3}(Z_{\text{PR}} + Z_{\text{TH}}^{\text{gated}} + Z_{\text{UV}}^{\text{gated}}) \in \mathbb{R}^{B\times C\times H/4\times W/4}. \tag{19}$$

This approach enables selective incorporation of auxiliary information based on primary feature guidance (Figure 9).

PL-SIG (pre-logit sigmoid gating). As a stand-alone, attention-free baseline (adapted from GF-Net [**?**]), we fuse once at the decoder's pre-logit tap (quarter resolution) using the primary pre-logit as the base Z_{PR}. For each auxiliary Z_i, a single-channel gate is predicted from the concatenation $[Z_{\text{PR}}|Z_i]$ via a two-layer 1×1 MLP (reduction r=4, ReLU, sigmoid), and the auxiliary is projected on the residual path and masked: $r_i = \psi_i \odot (W_i^{(t)} * Z_i)$. The fused pre-logit is $Z_{\text{fused}} = Z_{\text{PR}} + \sum_i r_i$, which a 1×1 classifier maps to logits before upsampling. Gates are computed from the current base. See Figure 10.

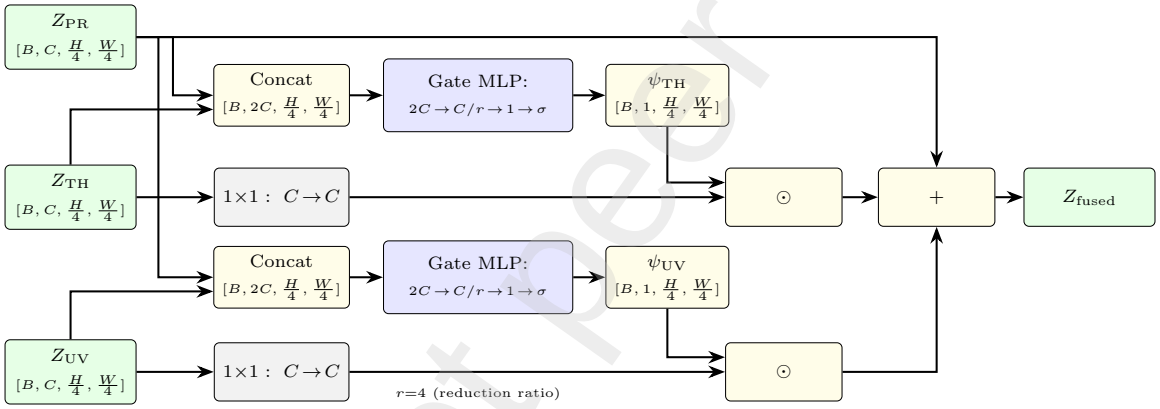


Figure 10: **Pre-logit sigmoid gating.** The base is the decoder's primary pre-logit Z_{PR}; each auxiliary Z_i is concatenated with Z_{PR} and passed through a two-layer 1×1 gating MLP (reduction r=4) to produce a single-channel spatial mask $\psi_i \in [0,1]$ at quarter resolution. Auxiliaries are projected on the residual path (if needed), masked, and added to form the fused pre-logit Z_{fused}, which is then classified.

Novel Implementations. We evaluate two proposed fusion mechanisms: CMAG (Section 4.3) and MWPA (Section 4.5).

Component Ablations. To isolate the individual contributions of CMAG's constituent mechanisms, we evaluate GCMA (Global Context Modality Attention) and sigmoid gating (PL-SIG) as standalone components. Both are described in detail in Section 4.3, with GCMA providing the cross-modal attention mechanism and SIG providing per-pixel spatial refinement. These ablations quantify the contribution of each component to CMAG's overall performance.

Training Protocol. All methods are trained end-to-end using AdamW optimisation, with a batch size of 6 for 220 epochs. The MiT-B0 backbone employs three independent decoder heads (primary RGB+DIN, thermal, UV), each with per-modality supervision using CEDice loss, manual class weights, and rare-class oversampling. Per-head learning rates are 9×10^{-4} for UV and thermal, and 1.1×10^{-3} for primary, with cosine annealing and head-specific warmup schedules. Multi-head supervision applies loss weights α_{pr}=0.75, α_{th}=0.35, α_{uv}=0.35, and α_{fused}=1.0. Modality dropout (p=0.10) is applied to auxiliary streams during training.

Normalisation swap at evaluation. For all the pre-logit fusion modules, LN layers are replaced by GN-16 at test time (channels divisible by 16), carrying over the learnt (γ, β).

Table 2 summarises the adapted methods. Source code and trained models will be made publicly available.

Table 2: Decoder-level fusion methods adapted for three-modality (RGB+DIN, Thermal, UV) pre-logit integration. Methods with **PL-** prefix are our adaptations preserving core innovations within the unified framework. GCMA and SIG are CMAG's component mechanisms evaluated standalone.

Method	Core Mechanism	Key Innovation	Our Adaptation
PL-MMTM	Global squeeze-excite over concatenated channels	Cross-modal channel transfer via shared global context	Three pre-logits $[Z_{PR}\|\|Z_{TH}\|\|Z_{UV}]$ concatenated, SE (r=8), single classifier
PL-R2AU	Recurrent attention gates with primary as gating signal	Spatial attention masks modulate auxiliary contributions	Primary gates thermal/UV via attention gates (F_{int}=256); average fusion of gated features
MWPA	Modality-wise parallel channel and spatial attention	Per-modality attention weights for selective fusion	Concat pre-logits → channel attention (r=16) → spatial attention (3×3 conv) → modality-wise weighting → classifier
PL-SIG	Per-pixel sigmoid gating	Lightweight spatial masks without attention	Base feature + gated residuals from thermal/UV; r=4 reduction
GCMA	Global context cross-modal attention	Modality-level attention ($O(M^2)$ vs $O(N^2)$)	Pool → Stack (B×3×C) → Multi-head attn (h=4) → Broadcast → classifier
CMAG	GCMA + SIG	Global attention for correspondence + fine-grained spatial gating	GCMA context + sigmoid-gated residuals; unified training

5.2. Robustness Evaluation

Robustness to sensor degradation, geometric misalignment, and incomplete multimodal input is critical for deployment applications. We systematically evaluate architectural resilience across three perturbation categories under controlled degradation, spanning three RGB lighting conditions (underexposed/RGB1, optimal/RGB3, overexposed/RGB5).

Modality Dropout Analysis. To simulate sensor failure and assess graceful degradation, we systematically ablate individual modalities and measure the resultant impact on fusion performance. Each modality (RGB, DIN, thermal, UV) is independently removed at inference while maintaining all other inputs, testing each method's capacity to preserve functionality under incomplete multimodal input. This establishes the relative importance and contribution of each sensor stream within the fusion framework, quantifying how critically the architecture depends on each modality.

Noise Resilience Testing. We evaluate performance degradation under four noise types—Gaussian, salt-and-pepper, speckle, and generic additive noise—applied independently to individual modalities (RGB, DIN, thermal, and UV). Noise intensity is varied across fourteen levels, with fine-grained sampling at low intensities (0.1, 0.2, 0.3, 0.4, 0.5) transitioning to coarser increments at higher intensities (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0), capturing both subtle corruption and severe degradation regimes. Figure 11 illustrates representative noise perturbations at three intensity levels across all modalities, demonstrating the range of corruption severity evaluated. This design yields 224 noise configurations (4 types, × 14 intensities, × 4 modalities), evaluated across three lighting conditions, resulting in 672 scenarios per method. With six methods under comparison (CMAG, GCMA, MWPA, PL-MMTM, PL-R2AU, PL-SIG), this totals 4,032 noise robustness evaluations, providing a comprehensive characterisation of architectural sensitivity to sensor corruption.

Spatial Misalignment Testing. We assess robustness to geometric misregistration by applying controlled spatial shifts to thermal and UV modalities, simulating realistic sensor calibration drift or mechanical misalignment. Each auxiliary modality is independently displaced by 20 and 40 pixels in the four cardinal directions (up, down, left, right), while the primary RGB-DIN stream remains stationary as the reference coordinate frame. Figure 12 visualises the effect of spatial misalignment on thermal and UV inputs, illustrating how features become spatially inconsistent with the primary stream. This produces 16 shift configurations (2 distances × 4 directions × 2 modalities), evaluated across three lighting conditions, yielding 48 misalignment scenarios per method. Across six methods, this generates 288 spatial robustness evaluations. These test-time shifts evaluate each method's intrinsic capacity to maintain performance under geometric inconsistency without calibration or retraining.

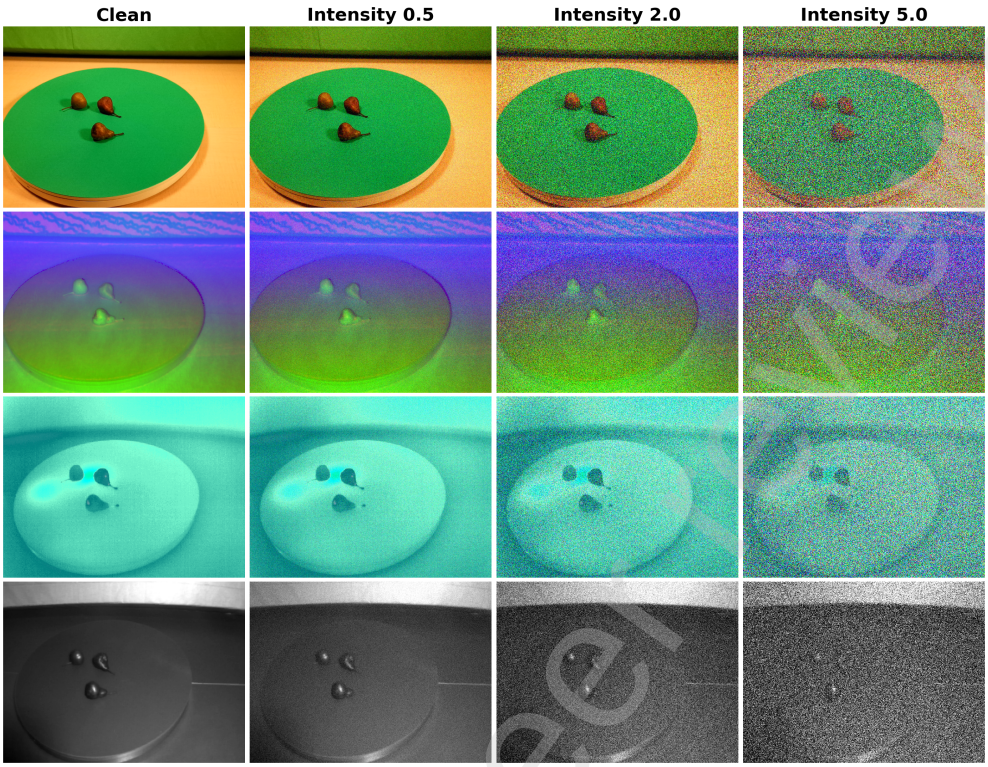


Figure 11: Representative noise perturbations applied during robustness evaluation. Rows show different modalities (RGB, DIN, thermal, UV); columns show clean input and three noise intensity levels (0.5, 2.0, 5.0) for Gaussian noise. At intensity 0.5, corruption is subtle; at 2.0, significant degradation is visible; at 5.0, severe corruption challenges recognition. Similar patterns apply to salt-and-pepper, speckle, and uniform noise types (not shown). All RGB examples are from the RGB3 (optimal lighting) test set.

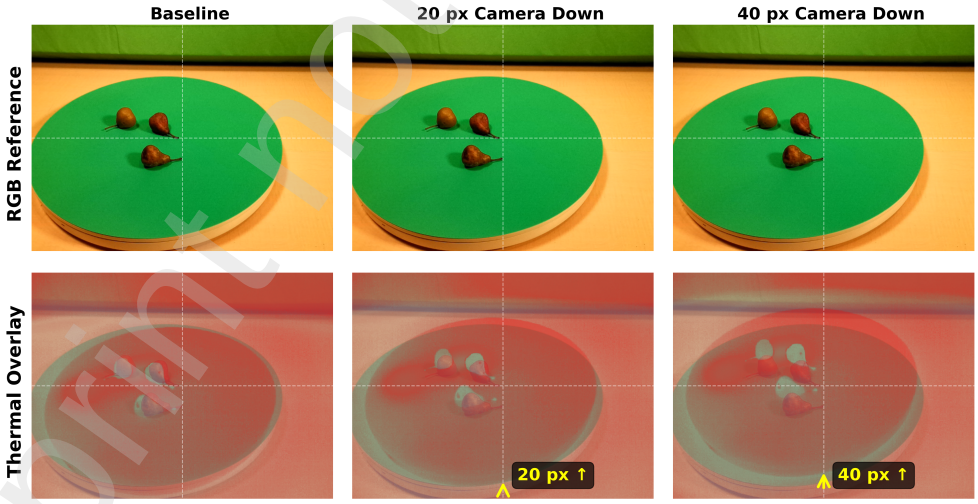


Figure 12: Spatial misalignment examples for thermal modality. Left: aligned baseline showing RGB and thermal overlay with default trained spatial registration. Middle: 20-pixel upward shift of thermal modality, creating moderate misregistration. Right: 40-pixel upward shift of thermal modality, demonstrating severe geometric inconsistency. Coloured overlays highlight spatial discrepancies between primary and auxiliary features. The RGB image remains fixed as the reference coordinate frame throughout all shift scenarios.

Impact Metrics. For a given RGB lighting condition r and perturbation scenario a, we quantify performance degradation relative to the unperturbed baseline ($\text{mIoU}_r^{\text{Full}}$) using two complementary metrics:

$$\text{Impact}_{\%}(r,a) = 100\,\frac{\text{mIoU}_r^{\text{Full}} - \text{mIoU}(r,a)}{\text{mIoU}_r^{\text{Full}}}, \tag{20}$$

$$\text{Impact}_{\text{pp}}(r,a) = \text{mIoU}_r^{\text{Full}} - \text{mIoU}(r,a), \tag{21}$$

where $\text{Impact}_{\%}$ quantifies relative degradation (normalised by baseline performance) and $\text{Impact}_{\text{pp}}$ measures absolute loss in percentage points. We prioritise $\text{Impact}_{\text{pp}}$ for cross-method comparisons, as it avoids the baseline bias inherent in relative metrics; methods with lower baseline accuracy may appear artificially robust when evaluated via percentage degradation. Per-scenario results are aggregated as mean ± standard deviation across the full test split, with lighting-specific statistics reported separately to isolate illumination-dependent impacts.

6. Results and Discussion

We evaluate six decoder-level fusion architectures across three lighting conditions. Systematic robustness analysis quantifies performance under modality dropout, spatial misalignment, and noise corruption. Results are organised by baseline performance, ablation studies, and comparative discussions of fusion strategies.

6.1. Evaluation Metrics

We report five standard segmentation metrics for comprehensive evaluation. Our primary metric, mean Intersection over Union (mIoU), provides class-balanced accuracy by averaging IoU across all classes. Frequency-Weighted IoU (FIoU) emphasises prevalent classes by weighting performance by pixel frequency. Pixel-level metrics include mean pixel accuracy (MPA), which computes per-class recall without penalising false positives, and pixel accuracy (PA), which measures overall pixel-level correctness. For distribution-free comparison, we report Mean Rank [**?**], where methods are ranked by IoU within each class and ranks are averaged across classes (a lower value indicates superior performance).

6.2. Overall Performance

Table 3 presents comprehensive baseline performance across all six fusion architectures under three lighting conditions. CMAG achieves the highest average mIoU of 84.18% across lighting conditions, validating its hybrid fusion strategy, which combines global context attention with fine-grained spatial gating. Under optimal lighting (RGB3), both CMAG (87.61%) and GCMA (86.72%) achieve strong performance, with a modest 0.89 percentage point difference. However, CMAG demonstrates superior robustness under challenging illumination; while GCMA degrades substantially (RGB1: 80.49%, RGB5: 78.03%), CMAG maintains more stable performance (RGB1: 82.54%, RGB5: 82.38%), providing a 3.2 percentage point advantage that is 3.6-fold larger than the optimal-lighting gap.

The remaining architectures achieve competitive baseline accuracy, ranging from 81.43% (PL-MMTM) to 83.14% (PL-R2AU) average mIoU, with reduced computational overhead (19M parameters vs. 22M for CMAG/GCMA). PL-R2AU demonstrates consistent performance across lighting conditions (average: 83.14%), while MWPA achieves an average mIoU.of 82.06%. PL-MMTM shows moderate baseline accuracy (average: 81.43%) but demonstrates superior robustness under perturbations, as detailed in Section 6.7. The lightweight sigmoid gating baseline (PL-SIG) achieves 82.29% average mIoU with balanced performance across conditions.

Mean rank analysis reveals a consistent performance ordering. Under optimal lighting (RGB3), CMAG achieves the best mean rank (3.12), followed by GCMA (4.56), PL-R2AU (5.55), and MWPA (5.88). Under suboptimal lighting, CMAG maintains strong rankings (RGB1: 10.28, RGB5: 9.05), while GCMA exhibits greater rank sensitivity to lighting degradation (RGB1: 12.59, RGB5: 12.12). All architectures sustain real-time throughput (31–34 FPS); detailed computational costs and throughput are reported in Section 6.6.

Frequency-weighted IoU (FIoU) is near-ceiling across all methods (typically 99.3–99.6%), reflecting class imbalance dominated by background and large, frequent items. The resulting mIoU–FIoU gap (about 12–16 pp across methods) indicates that errors concentrate in rarer or visually ambiguous classes; we analyse these in Section 6.5.

Table 3: Network comparison across lighting conditions and fusion architectures. Overall metrics (mIoU, FIoU, MPA, PA) and mean rank(1–18, lower is better) scores are reported across the three light settings of the six fusion methods. Best value per RGB configuration in bold. RGB1: underexposed; RGB3: optimal; RGB5: overexposed lighting.

Metric	RGB1 CMAG	RGB3 CMAG	RGB5 CMAG	RGB1 GCMA	RGB3 GCMA	RGB5 GCMA	RGB1 PL-R2AU	RGB3 PL-R2AU	RGB5 PL-R2AU	RGB1 MWPA	RGB3 MWPA	RGB5 MWPA	RGB1 PL-SIG	RGB3 PL-SIG	RGB5 PL-SIG	RGB1 PL-MMTM	RGB3 PL-MMTM	RGB5 PL-MMTM
Mean Rank	**10.3**	**3.1**	**9**	12.6	4.6	12.1	11.8	5.6	10.5	13.5	5.9	9.9	10.9	6.2	10.9	12.6	8.5	13.1
mIoU (%)	**82.54**	**87.61**	**82.38**	80.49	86.72	78.03	81.39	86.02	82.02	78.99	85.71	81.47	81.27	85.31	80.29	80.76	84.09	79.45
FIoU (%)	**99.37**	**99.55**	**99.37**	99.29	99.53	99.28	99.30	99.51	99.35	99.26	99.50	99.36	99.33	99.48	99.32	99.27	99.46	99.28
MPA (%)	**89.21**	**92.58**	**88.98**	88.00	91.87	85.18	88.13	91.86	88.21	86.21	91.70	87.30	88.75	91.45	87.04	87.76	90.17	86.54
PA (%)	**99.66**	**99.76**	**99.66**	99.61	99.75	99.61	99.62	99.74	99.65	99.60	99.73	99.65	99.64	99.72	99.63	99.61	99.70	99.61
FPS	31	31	31	31	31	31	34	34	34	32	32	32	34	34	34	34	34	34
Params (M)	22	22	22	22	22	22	19	19	19	19	19	19	19	19	19	19	19	19
GFLOPs	91.7	91.7	91.7	89.2	89.2	89.2	79.0	79.0	79.0	84.0	84.0	84.0	84.0	84.0	84.0	74.0	74.0	74.0

6.3. Normalisation Strategy: LayerNorm vs GroupNorm Trade-offs

Table 4 compares the impact of normalisation choice on accuracy and throughput across all six fusion architectures. LayerNorm (LN) during training consistently yields the highest mIoU across methods, establishing it as the preferred normalisation for learning decoder-level fusion. However, LN incurs substantial computational cost at inference: on C=512 feature maps at $\frac{H}{4}\times\frac{W}{4}$ = 120×160 resolution, LN achieves only 13.5–14.1 FPS across methods.

Switching to GroupNorm (GN) at inference while retaining LN-trained weights provides a favourable accuracy-throughput trade-off. GN-16 maintains accuracy within 0.2 percentage points of LN for most methods (CMAG: 87.61% → 87.45%, GCMA: 86.84% → 86.72%), while improving throughput by approximately 2.4 times. Finer grouping (GN-8) offers minimal accuracy improvement over GN-16 at reduced throughput, while coarser grouping (GN-32) marginally increases speed but exhibits slight accuracy degradation for some methods (MWPA: 85.71% → 85.35%, PL-SIG: 85.31% → 85.24%).

Training with GroupNorm directly (rather than LN) was briefly explored but yielded inferior performance, suggesting that global normalisation statistics during learning are important for cross-modal attention mechanisms. The asymmetric LN-train/GN-16 inference scheme, therefore, represents the optimal configuration, balancing accuracy with real-time inference requirements.

Table 4: Normalisation method comparison under optimal lighting (RGB3, 220 epochs). LayerNorm (LN) used during training for all configurations; GroupNorm (GN) variants applied at inference only. Best mIoU per method in bold. All methods show 2.3–2.4× throughput improvement with GN-16 inference while maintaining accuracy within 0.2pp of LN.

Method	LN (Inference)				GN-8				GN-16				GN-32			
	mIoU	FIoU	MPA	FPS	mIoU	FIoU	MPA	FPS	mIoU	FIoU	MPA	FPS	mIoU	FIoU	MPA	FPS
CMAG	**87.61**	99.55	92.58	13.51	87.43	99.54	92.69	29.90	87.45	99.54	92.52	31.09	87.42	99.53	92.38	31.32
GCMA	**86.84**	99.54	91.89	13.78	86.77	99.54	91.83	30.93	86.72	99.53	91.87	32.06	86.68	99.53	92.02	32.64
MWPA	85.67	99.51	90.71	13.92	**85.72**	99.51	91.16	31.80	85.71	99.50	91.70	33.01	85.35	99.47	92.21	33.32
PL-SIG	**85.59**	99.50	91.03	13.97	85.56	99.50	91.23	31.98	85.31	99.48	91.45	33.20	85.24	99.46	91.78	33.83
PL-MMTM	84.05	99.46	89.51	14.07	84.07	99.46	89.80	32.54	**84.09**	99.46	90.17	33.80	83.78	99.41	91.01	34.49
PL-R2AU	**86.07**	99.53	91.38	13.98	86.00	99.52	91.65	32.26	86.02	99.51	91.86	33.49	85.90	99.50	92.17	34.19

6.4. Lighting Condition Sensitivity

RGB illumination quality substantially affects segmentation performance across all decoder-level fusion architectures. We evaluate three illumination settings (RGB1: underexposed, RGB3: optimal, RGB5: overexposed) with results in Tables 3 and B.12.

Performance Across Lighting Conditions. Under optimal illumination (RGB3), methods achieve peak performance with narrow differentiation: CMAG (87.61% mIoU), GCMA (86.72%), PL-R2AU (86.02%), MWPA (85.71%), PL-SIG (85.31%), and PL-MMTM (84.09%) — a spread of only 3.52 pp. Underexposure (RGB1) shows a similar spread of 3.55 pp, with CMAG maintaining 82.54% (5.07 pp loss) and MWPA degrading to 78.99% (6.72 pp loss). Overexposure (RGB5) expands the spread to 4.35 pp, with GCMA suffering the largest degradation to 78.03% (8.69 pp

loss). Averaging across methods, RGB3→RGB5 degradation (5.30 pp) slightly exceeds RGB3→RGB1 (5.00 pp), reflecting irreversible saturation-induced information loss that auxiliary modalities cannot fully recover.
Mean rank analysis reveals differences in architectural stability: CMAG maintains consistent rankings (RGB3: 3.12, RGB1: 10.28, RGB5: 9.05), whereas GCMA exhibits high volatility (RGB3: 4.56, RGB1: 12.59, RGB5: 12.12), demonstrating a strong dependence on RGB quality.

Surface-Dependent Vulnerability. Class-level analysis reveals that surface properties dominate lighting robustness, far exceeding architectural differences. Reflective objects distinguished primarily by colour and pattern rather than geometric shape exhibit severe overexposure sensitivity: Apple degrades 27.50 pp (96.72%→69.22% at RGB5) for CMAG, as saturation obliterates the discriminative colour and patterns. In contrast, geometrically similar but less reflective Apple Green loses only 1.79 pp (95.80%→94.01%). Texture-rich objects maintain robust performance across lighting extremes: Grapes Blue achieves stable segmentation (RGB1: 94.24%, RGB3: 94.38%, RGB5: 95.92%), as geometric texture features survive both underexposure and saturation, where colour and specularity fail.
Sophisticated fusion mechanisms provide measurable advantages when auxiliary modalities are essential for discrimination. Under optimal lighting (RGB3), Carrot and Carrot Fake are visually nearly identical in RGB-DIN; yet, thermal signatures differ markedly (plastic vs. organic emissivity). Here, the architectural capacity to leverage thermal cues determines performance: CMAG achieves 90.05% on Carrot and 79.74% on Carrot Fake, while GCMA achieves 84.89% and 59.38% respectively, a 20.36 percentage point gap in replica discrimination performance. This demonstrates that CMAG's hybrid attention-gating mechanism more effectively weights thermal features when visual appearance alone is insufficient. Under underexposure (RGB1), where RGB colour and intensity cues are severely degraded while visual similarity between real and replica persists, PL-SIG excels on Carrot Fake (80.90%) through effective thermal gating, whereas PL-MMTM struggles (64.97%). This pattern highlights that the architectural capacity to modulate auxiliary contributions becomes critical when primary RGB features degrade, forcing greater reliance on thermal discrimination.
However, no architecture overcomes severe material-specific failures when primary modality features collapse; reflective Apple under overexposure saturation (27.50 pp loss) and Onion Red similarly degraded by saturation (11.62 pp loss) demonstrate fundamental limits where auxiliary modalities cannot compensate for irreversible information loss in the primary stream. Lighting robustness is therefore jointly determined by the architectural capacity to leverage auxiliary modalities when discriminative, the material properties governing primary feature preservation under lighting extremes, and the information-theoretic limits when primary features are irreversibly lost to sensor saturation or severe underexposure.

6.5. Challenging Classes and Modality-Specific Contributions

Background and container objects achieve near-ceiling accuracy across methods (Background: 99.77-99.86%, Bowl: 89.35-93.22%), whereas several categories expose persistent multimodal fusion challenges.

Partially Decayed Fruit. Partially rotten items (Apple Green Bad, Lemon Bad, Mandarin Bad) span a wide range of difficulty (49.74–94.26% IoU envelope across methods and lighting). Under optimal lighting (RGB3), Apple Green Bad exceeds 90% IoU for all six methods (range: 90.60-94.26%). Yet, performance separates dramatically under suboptimal conditions: at RGB1, the range widens to 60.01-72.17%, with PL-SIG being the most stable (72.17%/92.93%/73.07% at RGB1/3/5) and MWPA showing the most extensive spread (60.01%/94.26%/75.39%). For Mandarin Bad, RGB5 proves particularly fragile (GCMA: 23.14%), highlighting severe exposure sensitivity on decayed surfaces. The thermal modality provides critical complementary information here, as decay alters surface temperature and emissivity patterns that remain discriminative when RGB features degrade. Networks with explicit gating mechanisms (CMAG, PL-SIG) maintain superior cross-lighting stability, suggesting effective learnt thermal utilisation.

Synthetic Replica Objects. Replica plastics (Carrot Fake, Apple Fake, Lemon Fake) prove consistently challenging, with class envelopes of 59.38-80.90% (Carrot Fake), 76.54-94.34% (Apple Fake), and 41.13-74.34% (Lemon Fake). On Carrot Fake, PL-SIG achieves 80.90% at RGB1, while GCMA struggles at RGB3 (59.38%), indicating that challenging material discrimination benefits from methods that modulate auxiliary contributions effectively. However, the optimal architecture varies with illumination. The difficulty stems from the near-identical visual appearance of organic counterparts. Plastic exhibits distinctly different thermal emission (lower emissivity, faster thermal equilibration) and UV fluorescence, making thermal and UV modalities essential for this discrimination task.

Temperature-Discriminable Objects. For thermally separable classes, cross-method spreads are small. Cup Hot spans 88.62-94.75% overall, with per-lighting ranges of RGB1: 88.62-94.46% (5.8 pp), RGB3: 92.59-94.98% (2.4 pp), RGB5: 91.54-93.88% (2.3 pp). Cup Cold spans 82.90-95.08% with RGB1: 82.90-92.12% (9.2 pp), RGB3: 92.47-95.06% (2.6 pp), RGB5: 90.94-95.08% (4.1 pp). Under optimal or overexposed lighting, method variation is $\lesssim$2.6 pp, while underexposure increases the spread to 5.8 to 9.2 pp. This pattern validates that thermal provides unambiguous discriminative features when temperature differences are pronounced.

Implications for Fusion Design. Difficult classes exhibit wide intra-class performance ranges: Carrot Fake varies by ∼22 pp and Apple Green Bad by ∼34 pp across methods and lighting, compared with ≤0.8 pp for Mirror (98.19-98.96% range). Methods with gating or attention-based fusion (CMAG, GCMA, PL-SIG) rank better in challenging categories under favourable lighting (RGB3 mean ranks: CMAG 3.12 vs. PL-MMTM 8.50, yet no single architecture dominates across all lighting regimes and classes. The 6-9× greater performance variation on ambiguous categories compared to simple objects demonstrates that the architectural capacity to selectively leverage auxiliary modalities critically determines performance when RGB cues are insufficient. However, exposure-induced failures at RGB5 can overwhelm auxiliary compensation for certain materials, establishing fundamental limits of decoder-level fusion regardless of architectural sophistication.

6.6. Computational Efficiency and Real-Time Performance

All six decoder-level fusion architectures sustain real-time throughput on a single NVIDIA RTX 3090 GPU at 640×480 resolution (Table 3). Parameter counts cluster in two tiers: 22M for CMAG and GCMA (attention-based methods), and 19M for MWPA, PL-R2AU, PL-SIG, and PL-MMTM (lighter fusion mechanisms). Computational cost spans 74.0 to 91.7 GFLOPs, with CMAG (91.7 GFLOPs) and GCMA (89.2 GFLOPs) representing the upper bound, MWPA (84.0 GFLOPs), PL-R2AU (79.0 GFLOPs), and PL-SIG (84.0 GFLOPs) occupying the middle range, and PL-MMTM (74.0 GFLOPs) offering the most efficient configuration. Despite these variations, measured throughput remains tightly bounded at 31-34 FPS across all methods, demonstrating that decoder-level fusion introduces minimal overhead beyond the shared encoder-decoder backbone.

The accuracy-efficiency trade-off favours adaptive fusion mechanisms. CMAG achieves the highest average mIoU of 84.18% (mean across RGB1/3/5) at 31 FPS with 91.7 GFLOPs. In comparison, the most efficient method, PL-MMTM, operates at 34 FPS with 74.0 GFLOPs while achieving an average mIoU of 81.43%. Relative to PL-MMTM, CMAG incurs a 24% computational overhead (91.7 vs. 74.0 GFLOPs) and an 8.8% throughput reduction in exchange for a 2.75 pp mIoU improvement. Other methods occupy intermediate positions: PL-R2AU (79.0 GFLOPs, 34 FPS, 83.14% mIoU) and MWPA (84.0 GFLOPs, 32 FPS, 82.06% mIoU) offer balanced alternatives, while GCMA (89.2 GFLOPs, 31 FPS, 81.75% mIoU) demonstrates that computational cost alone does not guarantee superior accuracy; architectural design and lighting robustness are critically important.

All configurations maintain throughput exceeding 30 FPS at VGA resolution, meeting real-time requirements for online robotic tasks. The modest computational differences between methods (17.7 GFLOPs range, 3 FPS variation) relative to substantial accuracy variations (2.75 pp between best and worst) underscore that the decoder-level fusion strategy, rather than raw computational capacity, determines segmentation quality in challenging multimodal scenarios.

6.7. Modality Importance: Ablation Studies

We quantify each modality's contribution and alignment sensitivity through two ablation families: (i) drop ablations (complete removal of a modality at inference) and (ii) spatial shift ablations. Unless stated otherwise, results are averaged across all six networks and three lighting conditions. Comprehensive class-wise robustness visualisations for all decoder-level architectures are provided in Appendix Appendix B.3.

6.7.1. Drop Ablation Results

Complete modality removal establishes a consistent importance hierarchy: RGB > DIN > T24 > U8. Figure 13(left) presents the consolidated results, with mean absolute mIoU losses averaged across networks and lighting conditions:

- **RGB drop:** 59.50 pp loss (72.05% relative degradation)
- **DIN drop:** 49.61 pp loss (60.53% relative degradation)
- **T24 drop:** 24.62 pp loss (29.79% relative degradation)

- **U8 drop:** 16.82 pp loss (20.36% relative degradation)

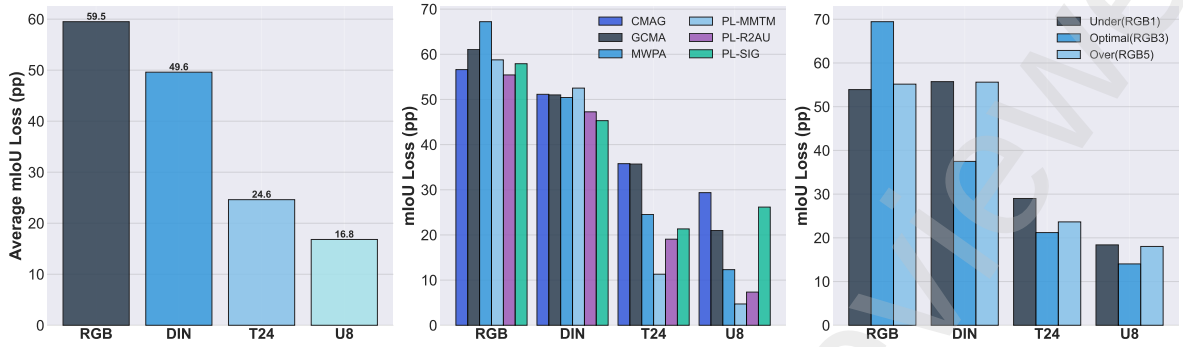


Figure 13: Overview of drop ablation impact. (left) Overall modality importance averaged across all networks and lighting conditions, ranked from most to least critical. (centre) Network comparison showing impact for all six fusion networks across the four modalities, revealing architecture-specific dependencies. (right) RGB variant comparison showing how lighting conditions (underexposed, optimal, overexposed) affect modality importance when averaged across networks.

The dominance of RGB and DIN reflects their role as the primary feature extractors for geometric structure and appearance, while thermal and UV provide complementary discriminative cues for challenging material classes (Section 6.5). Table 5 presents per-network residual mIoU after each modality is dropped, revealing substantial architectural variation: CMAG and GCMA exhibit the highest sensitivity (mean losses of 43.22 pp and 42.19 pp, respectively), while PL-MMTM and PL-R2AU demonstrate superior resilience (31.82 pp and 32.27 pp).
One illustrative example demonstrates the magnitude of primary modality dependence: CMAG under optimal lighting (RGB3) degrades from 87.61% mIoU baseline to 17.49% when RGB is removed (a 70.12 pp loss), confirming the critical role of the primary visual stream.

Table 5: Drop ablation results averaged over RGB scenarios: residual mIoU (%) after removing each modality. Networks ordered by robustness (left to right: least robust to most robust). Lower values indicate greater modality dependence. The bottom row shows the mean loss.

Dropped	**RGB**	**CMAG**	**GCMA**	**MWPA**	**PL-SIG**	**PL-R2AU**	**PL-MMTM**
RGB	RGB1	34.51	30.09	11.81	37.19	31.09	19.75
	RGB3	17.49	11.73	11.81	11.91	19.41	21.80
	RGB5	30.70	20.15	23.38	25.04	32.65	21.80
DIN	RGB1	18.54	24.01	27.63	27.89	33.40	22.05
	RGB3	57.86	45.37	42.10	56.16	52.26	32.17
	RGB5	22.73	22.86	27.63	27.89	21.98	27.91
T24	RGB1	36.30	47.97	62.48	55.85	48.07	63.23
	RGB3	61.99	31.51	65.74	72.15	77.61	74.60
	RGB5	46.89	58.68	46.83	55.85	66.57	67.92
U8	RGB1	55.48	56.48	70.57	50.03	74.44	70.61
	RGB3	53.68	70.79	70.57	69.31	82.05	80.21
	RGB5	55.27	55.00	70.57	50.03	70.89	74.70
Mean loss (pp)		43.22	42.19	38.63	37.68	32.27	31.82

6.7.2. Spatial Shift Ablations

We assess robustness to sensor misalignment by applying controlled spatial shifts to thermal and UV images (20 px/40 px), while maintaining RGB-DIN as the reference coordinate frame, with results aggregated across shift configurations, networks, and lighting conditions.

Overall Impact. Spatial misalignment induces substantially lower degradation than complete modality removal (mean: 4.20 pp vs. 37.64 pp for drops), confirming that decoder-level fusion exhibits intrinsic tolerance to moderate sensor misregistration. Table 6 presents detailed per-network degradation under thermal and UV shifts. Degradation scales non-linearly with magnitude: 20 px shifts cause a 2.61 pp mean loss, while 40 px shifts induce a 5.68 pp loss, a 2.18×

increase. This super-linear relationship reflects that larger misalignments increasingly violate the spatial correspondence assumptions implicit in learnt fusion weights. The proportion of fusion failures (where multimodal performance falls below single-modality baselines) increases 3.6-fold from 8.2% at 20 px to 29.3% at 40 px, establishing a critical misalignment threshold beyond which auxiliary information degrades rather than enhances segmentation accuracy.

Table 6: Performance degradation (pp) under spatial misalignment of auxiliary modalities. Values show mean mIoU loss ± standard deviation when thermal or UV streams are shifted relative to the aligned RGB-DIN primary. All networks demonstrate moderate degradation at 20 px shifts (1.07–3.95 pp), which approximately doubles at 40 px shifts (2.71–8.40 pp), confirming the inherent tolerance of decoder-level fusion to geometric misalignment.

Network	**Thermal Shift**		**UV Shift**		**Mean Degradation**	
	20 px	40 px	20 px	40 px	20 px	40 px
CMAG	3.50±0.77	7.99±1.57	1.61±0.89	4.73±1.76	2.56	6.36
GCMA	3.05±0.86	6.92±1.54	1.07±0.73	3.06±1.64	2.06	4.99
MWPA	3.61±0.94	7.86±1.74	1.43±0.78	3.05±1.31	2.52	5.46
PL-MMTM	3.44±0.87	7.35±1.64	1.22±0.45	2.71±0.97	2.33	5.03
PL-R2AU	3.86±0.63	8.40±1.72	1.60±0.85	3.82±1.17	2.73	6.11
PL-SIG	3.95±0.68	7.49±2.92	3.00±1.40	4.72±1.24	3.48	6.11
Average	3.57±0.33	7.67±0.51	1.66±0.71	3.68±0.89	2.61	5.68

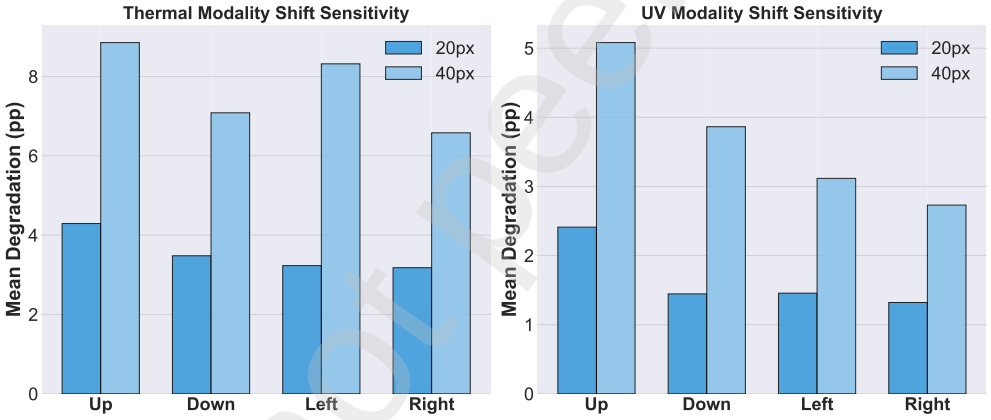


Figure 14: Spatial shift sensitivity by magnitude (20 px/40 px) and direction (up, down, left, right) for thermal and UV modalities, averaged across all networks and lighting conditions. Thermal exhibits mild vertical bias (upward shifts most damaging: 7.45 pp at 40 px) and consistently 1.63× greater sensitivity than UV across all scenarios. Degradation scales super-linearly with offset magnitude, doubling the shift produces 1.93× greater loss, indicating that geometric inconsistency tolerance degrades non-linearly beyond modest misalignments.

Modality-Specific Sensitivity. As shown in Table 6, thermal exhibits consistently greater sensitivity to misalignment than UV across all networks (mean degradation: thermal 3.57 pp at 20 px and 7.67 pp at 40 px vs. UV 1.66 pp at 20 px and 3.68 pp at 40 px). This differential reflects thermal's greater contribution to fusion performance, as demonstrated by drop ablations (Section 6.7); removing thermal causes a 24.62 pp mean loss compared to UV's 16.82 pp loss. Modalities with larger fusion contributions exhibit proportionally greater sensitivity to spatial misalignment, as misregistration directly degrades the discriminative features that the network has learnt to rely upon.

Directional Asymmetry. Thermal demonstrates a mild vertical bias, with upward shifts proving most damaging (20 px: 3.89 pp, 40 px: 7.45 pp), while UV exhibits near-isotropic behaviour (directional spread <0.6 pp at each magnitude), as shown in Figure 14. This divergence reflects the different object categories that contribute discriminative cues for each modality. Thermal-discriminable objects (temperature-dependent items) and UV-discriminable objects (material-dependent items) occupy different spatial distributions, sizes, and positions within the scene. The thermal vertical bias likely reflects that temperature-based objects in the dataset exhibit systematic vertical positioning patterns, while

UV-critical discrimination tasks (replica vs. organic material) occur across more spatially diverse object orientations. This pattern persists across all networks and lighting conditions, confirming that it originates from the dataset object distribution rather than architectural characteristics.

Architectural Robustness. Figure 15 presents comprehensive per-network robustness across all shift scenarios. The left heatmap shows mean fusion degradation (pp) for each network-scenario combination, while the right heatmap displays the percentage of scenarios where fusion remains beneficial (shifted multimodal mIoU exceeds single-modality baseline). A score of 100% indicates that fusion provides a positive benefit in all evaluated scenarios (across lighting conditions and directions) for that modality-magnitude combination, while lower percentages reveal cases where spatial misalignment causes fusion to underperform relative to single-modality baselines.

Figure 15: Network performance heatmap for spatial shift robustness. Each cell aggregates performance across 12 scenarios: three lighting conditions (RGB1/3/5) and four shift directions (up/down/left/right) for thermal/UV shifts at 20 px or 40 px magnitudes. **Left:** Mean fusion degradation (pp) shows performance loss—lower values (yellow) indicate robustness, higher values (red) indicate sensitivity. **Right:** Percentage of scenarios maintaining positive fusion benefit (shifted multimodal mIoU > single-modality baseline)—100% (green) indicates fusion remains beneficial across all conditions, while lower percentages (yellow/red) reveal cases where misalignment causes fusion to underperform.

Network robustness varies substantially: PL-SIG achieves 94% positive scenarios across all shifts, followed by GCMA (89%), CMAG (83%), PL-R2AU (80%), MWPA (72%), and PL-MMTM (70%). Notably, this ranking differs from baseline accuracy ordering (Table 3), demonstrating that peak performance and shift robustness are partially orthogonal properties. Simple gating mechanisms (PL-SIG) prove most tolerant to misalignment, while sophisticated fusion (CMAG) trades shift robustness for higher clean-data accuracy. The heatmap reveals that UV shifts at 20 px maintain near-perfect fusion benefit (>95% positive) across most networks, while thermal 40 px shifts prove challenging, with positive rates dropping to 50-75% for less robust architectures.

Lighting Interaction. Shift impact remains broadly stable across illumination conditions: mean degradations are 4.58 pp (RGB1), 3.87 pp (RGB3), and 3.90 pp (RGB5). The marginally higher sensitivity under underexposure (RGB1: 4.58 pp) reflects a benefit-fragility trade-off: while auxiliary modalities provide the most significant value when primary features degrade, they simultaneously become more critical to spatial correspondence, amplifying the impact of misalignment.

6.7.3. Summary

Complete modality removal represents the primary failure mode for decoder-level fusion, with RGB and DIN drops causing severe degradation (59.50 pp and 49.61 pp, respectively). In contrast, moderate spatial misalignment (20 px) induces minimal performance loss (2.87 pp average), validating that decoder-level fusion accommodates sensor misregistration without explicit alignment mechanisms. UV exhibits 2.4× greater spatial tolerance than thermal (3.20 pp vs. 5.20 pp mean degradation), correlating with its lower overall contribution to fusion (16.82 pp vs. 24.62 pp drop

impact). Architectural robustness to perturbations (PL-MMTM, PL-R2AU superior for drops; PL-SIG superior for shifts) does not correlate with baseline accuracy (CMAG, GCMA superior), revealing an inherent trade-off between peak performance and perturbation resilience.

6.8. Noise Robustness Analysis

We evaluate robustness to input corruption by applying four noise types (Gaussian, salt-and-pepper, speckle, and generic additive noise) independently to each of the four modalities (RGB, DIN, T24, U8) at 14 intensity levels (0.1-5.0). This yields 16 corruption scenarios evaluated across six networks and three lighting conditions, totalling 4,032 evaluations. Results quantify both absolute performance loss (percentage points, pp) and relative degradation (%) to characterise architectural resilience to sensor noise. Class-level vulnerability patterns under noise perturbations are visualised in Appendix Appendix B.3.

6.8.1. Overall Noise Sensitivity

Table 7 reports the mean loss across all modalities, noise types, intensities, and lighting conditions. PL-R2AU demonstrates the highest noise tolerance (8.80 pp mean loss, 10.6% relative degradation), followed by PL-MMTM (9.67 pp, 11.9%), with PL-SIG and GCMA occupying mid-tier positions. CMAG and MWPA exhibit the highest sensitivity (12.93 pp and 13.91 pp, respectively). Because degradation is non-linear with severity, we report anchor losses at both mild and severe levels of corruption. Between i=1.0 and i=5.0, degradation increases substantially: PL-R2AU from 6.51 to 16.74 pp (+10.23), CMAG from 8.77 to 25.59 pp (+16.82), and MWPA from 9.49 to 27.04 pp (+17.55). The widening performance spread under severe corruption confirms that architectural capacity to leverage auxiliary modalities determines robustness. Degradation is non-linear: at intensity 1.0, losses span 6.51 to 9.49 pp with modest inter-network variation (2.98 pp range); at intensity 5.0, this expands to 16.74 to 27.04 pp (10.30 pp range), demonstrating that architectural differences amplify 3.5× under severe corruption.

Table 7: Noise robustness summary across all modalities, noise types, and lighting. Anchor losses at mild (i=1.0) and severe (i=5.0) corruption with their difference. Lower is better.

Network	**PL-R2AU**	**PL-MMTM**	**PL-SIG**	**GCMA**	**CMAG**	**MWPA**
Mean loss (pp)	**8.80**	9.67	10.66	10.70	12.93	13.91
Relative deg. (%)	**10.6**	11.9	13.0	13.2	15.4	17.1
Mean loss @ i=1.0 (pp)	6.51	7.57	7.96	8.05	8.77	9.49
Mean loss @ i=5.0 (pp)	16.74	18.89	20.63	21.41	25.59	27.04
Δ (5.0–1.0) (pp)	10.23	11.32	12.67	13.36	16.82	17.55

Figure 16 presents degradation trajectories across all networks: an overall summary (left) and four modality-resolved panels (RGB, DIN, T24, U8) for direct comparison. PL-R2AU maintains the lowest or near-lowest curve across all panels, while MWPA and CMAG rise most steeply at high severities, driven primarily by their sensitivity to the UV modality U8. At mild corruption (intensity 1.0), losses span 6.51-9.49 pp (7.9-11.7% relative); at severe corruption (intensity 5.0), they reach 16.74-27.04 pp (20.2-33.1% relative).

6.8.2. Modality-Specific Vulnerability

Modality vulnerability follows a clear hierarchy. Table 8 quantifies per-modality mean losses across networks. RGB corruption proves most damaging (20.74-25.60 pp), followed by thermal (7.94-10.59 pp), UV (2.30-14.85 pp), and depth-intensity (2.78-5.02 pp). The worst single scenario is Gaussian noise on RGB, resulting in a 28.84 pp mean loss (35.0% relative degradation) across all networks. DIN exhibits the lowest sensitivity, reflecting its auxiliary role in providing geometric cues that remain largely intact under photometric noise. UV exhibits substantial network-dependent variation: while most architectures show low UV sensitivity (2.30-5.60 pp), CMAG and MWPA demonstrate anomalous vulnerability (13.96 and 14.85 pp, respectively).

Figure 17 presents scenario-level sensitivity across all noise types and intensities, confirming that RGB corruption dominates degradation, while DIN proves to be the most robust across architectures.

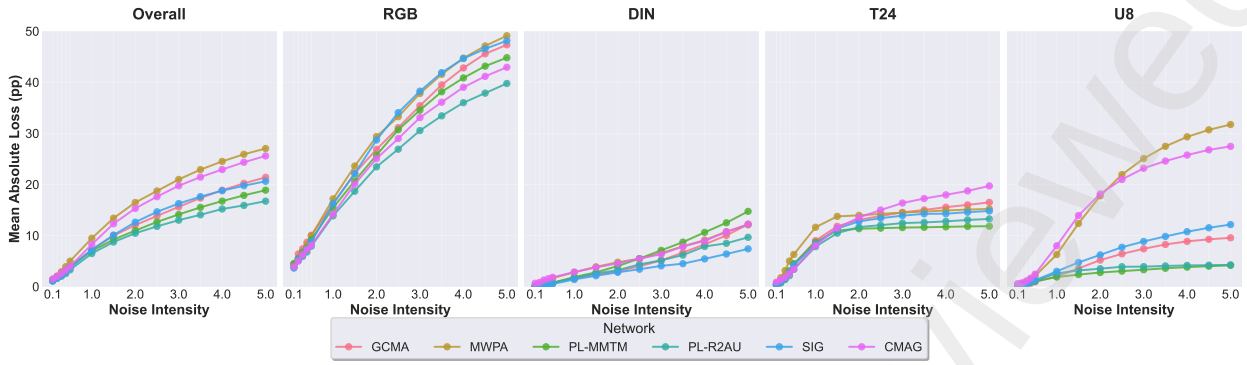


Figure 16: Network robustness to noise corruption across intensity levels (0.1-5.0). On the left, the overall loss averaged over all modalities, noise types, and lighting conditions. This is followed by modality-resolved panels (RGB, DIN, T24, U8) with shared y-axis (0-50 pp), enabling direct cross-modality comparison.

Table 8: Modality-specific mean loss (pp) by network, averaged over noise types and intensities. Bold indicates best performance per modality (lowest loss).

Modality	PL-R2AU	PL-MMTM	PL-SIG	GCMA	CMAG	MWPA
DIN	3.66	5.02	**2.78**	3.96	4.94	4.84
T24	8.16	**7.94**	9.14	9.75	10.59	10.35
U8	2.63	**2.30**	5.60	4.51	13.96	14.85
RGB	**20.74**	23.41	25.14	24.58	22.23	25.60

UV-Driven Vulnerability of CMAG and MWPA. CMAG and MWPA demonstrate disproportionate sensitivity to UV (U8) corruption compared to other networks, as evident in the U8 panel of Figure 16. CMAG suffers a 13.96 pp mean loss under U8 noise, compared to 2.30 to 5.60 pp for PL-MMTM, PL-R2AU, and PL-SIG; MWPA similarly degrades by 14.85 pp. This UV-specific vulnerability stems from CMAG's architectural design: its pre-logit attention-gating mechanism leverages UV cues more aggressively than other methods, causing unfiltered UV noise to propagate with high gain at severe intensities.

6.8.3. Architectural Patterns

Network rankings for noise robustness differ substantially from both baseline accuracy and spatial shift robustness, revealing distinct architectural trade-offs. PL-R2AU achieves the highest noise tolerance (8.80 pp mean degradation) despite moderate baseline performance (83.14% mIoU). Conversely, CMAG attains the highest baseline accuracy (84.18% mIoU) but exhibits greater noise sensitivity (12.93 pp mean degradation, 1.47× higher than PL-R2AU). This UV-driven vulnerability (13.96 pp loss under U8 corruption versus 2.30 to 5.60 pp for other networks) reflects CMAG's learnt fusion strategy: its hybrid attention-gating mechanism optimises auxiliary modality utilisation for clean-data performance, amplifying degradation when those channels contain noise.

The relationship between fusion strategy and perturbation tolerance reveals a fundamental trade-off. CMAG achieves the highest baseline accuracy (84.18% average mIoU) by learning to aggressively exploit auxiliary modalities; however, it consequently exhibits the highest noise sensitivity (12.93 pp mean degradation). In contrast, PL-R2AU sacrifices 1.04 pp average baseline accuracy (83.14% mIoU) yet achieves substantially superior noise resilience (8.80 pp mean degradation)—a 4.13 pp robustness improvement representing a 4.0× return on the accuracy sacrifice.

PL-MMTM demonstrates balanced performance: second-best for noise (9.67 pp mean degradation) and best for drop robustness (31.82 pp mean loss, Table 5), yet worst for spatial shifts (70% positive scenarios). This pattern underscores orthogonal robustness dimensions. Squeeze-and-excitation-based channel gating (PL-MMTM) enables robust handling of corrupted inputs through conservative channel weights; however, it struggles when spatial correspondence is violated, as global pooling discards the geometric structure. Conversely, spatial attention mechanisms (PL-SIG: 94% positive shift scenarios) tolerate misalignment through pixel-wise gating but can amplify auxiliary contributions, propagating

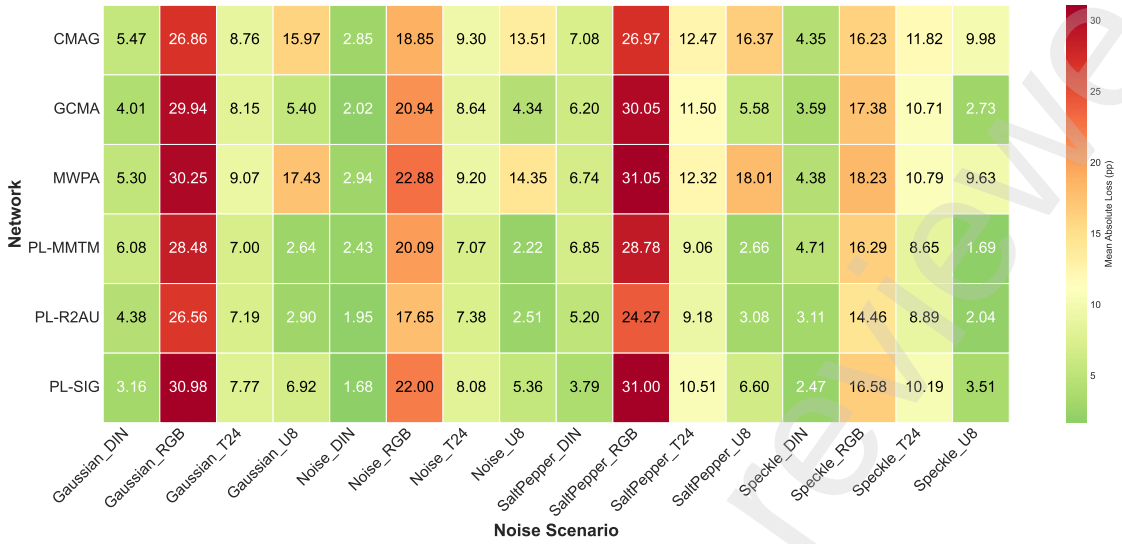


Figure 17: Scenario-level noise sensitivity heatmap (modality × noise type). Cell intensity represents mean absolute loss (pp) across networks, lighting conditions, and intensity levels. Gaussian noise on RGB constitutes the worst-case scenario (28.84 pp, 35.0% relative degradation). RGB-targeted corruption dominates across all noise types, while DIN and UV corruptions induce substantially lower degradation for most architectures.

noise through fusion pathways (10.66 pp mean degradation under noise).

6.8.4. Practical Implications

Architecture selection should account for both noise severity and modality-specific vulnerability. At mild corruption (intensity 1.0), all networks remain functional (6.51 to 9.49 pp loss, 7.9 to 11.7% relative degradation). At severe corruption (intensity 5.0), losses range from 16.74 to 27.04 pp, rendering some configurations effectively non-functional. Prioritising RGB sensor quality yields the most significant robustness gains, as RGB-targeted noise causes mean degradation of 20.74 to 25.60 pp, versus only 2.78 to 10.59 pp for auxiliary modalities—a 2.0 to 9.2× difference. For systems that heavily leverage UV features (CMAG, MWPA), UV channel quality control becomes critical to prevent severe degradation (13.96 to 14.85 pp under UV corruption, compared to 2.30 to 5.60 pp for UV-conservative methods). Optimal architecture choice depends on operational priorities and expected perturbation profiles. CMAG achieves the highest baseline accuracy (84.18% average mIoU) with acceptable noise tolerance at mild intensities (8.28 pp loss at intensity 1.0). However, under persistent moderate-to-severe noise (intensity > 1.5), PL-R2AU exhibits ∼35% lower degradation than CMAG at high severity (e.g., 16.74 vs. 25.59 pp at i=5.0), while PL-MMTM is ∼26–28% lower (e.g., 18.89 vs. 25.59 pp at i=5.0). When spatial misalignment is the dominant perturbation source, PL-SIG and GCMA achieve superior shift robustness (94% and 89% positive scenarios, respectively; Section 6.7.2), though with moderately higher noise sensitivity (10.66 pp and 10.70 pp, respectively).

6.9. Comparative Analysis of Fusion Strategies

Baseline evaluation across six fusion architectures reveals distinct performance-robustness trade-offs under varied operational conditions. CMAG achieves the highest average accuracy with stable lighting tolerance, while adapted methods demonstrate competitive performance with reduced computational overhead. Systematic ablations reveal orthogonal robustness dimensions that inform deployment strategy:

- **Modality importance is architecturally invariant.** Drop experiments establish a consistent hierarchy (RGB > DIN > T24 > U8) across all methods, indicating that information content dominates architectural effects. However, fusion strategies exhibit substantial variation in sensitivity: PL-MMTM achieves superior drop robustness through modality isolation, while CMAG's cross-modal integration amplifies dependence on auxiliary channels (Section 6.7.2).
- **Fusion aggressiveness determines accuracy-robustness trade-offs.** CMAG's aggressive auxiliary utilisation yields the highest baseline accuracy (84.18% mIoU) but the greatest noise sensitivity (12.93 pp), while PL-R2AU's

conservative fusion maintains competitive accuracy (83.14% mIoU) with superior robustness (8.80 pp). This pattern is most pronounced in UV corruption, where learnt fusion strategies determine vulnerability (Section 6.8).

- **Spatial robustness trades off against noise resilience.** Methods preserving spatial structure (PL-SIG, GCMA) tolerate geometric misalignment through pixel-wise correspondence, whereas global pooling architectures (PL-MMTM) discard spatial information for noise suppression. This reveals fundamental architectural constraints: spatial preservation enables alignment compensation but propagates corruption; global aggregation suppresses noise but eliminates geometric cues (Sections 6.7.2, 6.8).
- **Auxiliary modality utilisation determines lighting robustness.** CMAG's hybrid gating mechanism maintains stable performance under challenging illumination (RGB1: 82.54%, RGB5: 82.38%), outperforming GCMA by 3.2 percentage points under suboptimal lighting (RGB1/5: 80.49%/78.03%) through effective thermal and UV compensation when RGB features degrade. Underexposed conditions (RGB1) prove most challenging across all architectures, inducing 11% higher average degradation compared to optimal lighting (RGB3: 24.20 pp vs RGB1: 26.93 pp mean loss), confirming that challenging lighting conditions amplify dependence on auxiliary modalities.
- **Deployment context dictates optimal architecture.** For controlled environments prioritising peak accuracy, CMAG maximises performance with acceptable noise tolerance at mild intensities. Under persistent severe corruption, PL-R2AU and PL-MMTM exhibit substantially lower degradation through conservative fusion. For misalignment-dominated scenarios, PL-SIG and GCMA provide superior shift tolerance, with moderate noise sensitivity (Sections 6.7.2, 6.8).
- **Computational efficiency remains comparable across methods.** All architectures achieve real-time inference capability with modest overhead differences, positioning computational cost as a secondary selection criterion relative to accuracy-robustness trade-offs under expected perturbation profiles (Table 3).

6.10. Encoder- vs Decoder-Level Fusion

To contextualise decoder-level fusion within the broader multimodal segmentation landscape, we compare our pre-logit integration approach against encoder-level fusion, represented by GF-Net [**?**], our previous encoder-level architecture on the MM5 dataset. Both architectures employ MiT-B0 backbones and fuse RGB, depth-intensity-normals (DIN), thermal (T24), and UV (U8) modalities, enabling direct performance comparison while isolating the impact of fusion stage placement. Crucially, both networks employ Stage-Wise Intensity Fusion (SWIF) to enhance the RGB primary stream with DIN composites; the fundamental distinction lies in where auxiliary thermal and UV modalities are integrated within the feature hierarchy.

Architectural Paradigms. For direct architectural comparison, we focus on sigmoid gating methods that isolate fusion stage effects: GF-Net employs encoder-level sigmoid gating, while PL-SIG implements decoder-level sigmoid gating. This comparison enables us to hold fusion mechanism complexity constant while varying only the integration stage placement, thereby isolating the architectural impact of early versus late fusion.

Encoder-level fusion (GF-Net) applies SWIF to inject DIN into RGB features at each encoder stage (Stages 1–4), establishing an enhanced RGB, DIN primary representation. Thermal and UV auxiliaries are then fused into this primary stream at every encoder stage through per-pixel sigmoid gating following CM-FRM spatial alignment [**?**]. This stage-wise encoder fusion propagates multimodal features through all subsequent encoder depths and the shared decoder, enabling deep cross-modal interaction at the cost of tight architectural coupling and spatial alignment dependency.

Decoder-level fusion (PL-SIG) similarly employs SWIF to enhance RGB with DIN at each encoder stage within the primary stream. However, thermal and UV modalities are processed through independent encoder-decoder pipelines, as described in Sections 4.3.2 and 5.

Quantitative Performance Comparison. Table 9 presents accuracy and efficiency metrics for both paradigms. GF-Net achieves consistently higher baseline mIoU across all lighting conditions, with a 3.51 pp average advantage over PL-SIG (85.80% vs 82.29%). This accuracy premium is most pronounced under suboptimal illumination (RGB1: +3.63 pp, RGB5: +3.91 pp), suggesting that stage-wise encoder integration provides stronger illumination invariance through progressive refinement across encoder depths. Under optimal lighting (RGB3), the gap narrows to 2.99 pp, indicating that both paradigms achieve comparable performance when RGB quality is high. Amongst decoder methods,

Table 9: Encoder- vs decoder-level fusion at VGA resolution. GF-Net refers to the SWIF-Gated (RGB+DIN+T+UV) configuration; PL-SIG is the architecturally matched decoder method (sigmoid gating only); "Decoder (best)" is CMAG from Table 3; "Decoder (mean)" averages the six decoder-level methods. All methods use MiT-B0 backbones and SWIF-enhanced RGB+DIN primary streams on the same MM5 test split.

Method	RGB1	RGB3	RGB5	Mean	FPS	GFLOPs
GF-Net (encoder)	84.90	88.30	84.20	85.80	55	17.3
PL-SIG (decoder)	81.27	85.31	80.29	82.29	34	84.0
Decoder (best)	82.54	87.61	82.38	84.18	31	91.7
Decoder (mean)	80.99	85.91	80.99	82.63	31–34	74.0–91.7

CMAG achieves the highest average mIoU (84.18%), reducing the encoder advantage to 1.62 pp, though at an increased computational cost.
The efficiency disparity is substantial: GF-Net achieves 55 FPS vs PL-SIG's 34 FPS (62% higher throughput) while requiring only 17.3 GFLOPs vs 84.0 GFLOPs (79% lower computational cost). This 4.9× computational advantage stems from encoder fusion's single decoder pathway processing jointly refined features, as opposed to decoder fusion's independent per-modality encoder-decoder pipelines that operate until late integration. The parameter efficiency arises from GF-Net's shared decoder, which consumes fused encoder outputs, whereas PL-SIG maintains separate decoders for each modality stream.

Robustness Under Perturbation. Ablation studies reveal contrasting vulnerability profiles between sigmoid gating paradigms. RGB removal causes severe degradation across both approaches, yet decoder fusion retains substantially higher residual performance (PL-SIG: 24.71% mean residual mIoU) compared to encoder fusion (GF-Net: 14.05% mean residual mIoU), a 1.76× advantage, as shown in the drop columns of Figure 18. This resilience advantage demonstrates that modality isolation in decoder fusion, where thermal and UV maintain independent processing pathways until the pre-logit stage, enables more graceful degradation when the primary sensor fails. Conversely, encoder fusion's early integration creates representational dependencies that cannot be bypassed when the base modality is unavailable. Tables 10 and 11 quantify these differences systematically.

Table 10: Performance degradation (pp drop and relative degradation) under complete modality removal. Values show mean mIoU loss averaged across three lighting conditions (RGB1/3/5). Decoder methods demonstrate substantially lower degradation than encoder fusion, particularly for RGB drops where PL-SIG (57.91 pp) outperforms GF-Net (71.47 pp) by 13.56 pp. The decoder mean represents the average across all six decoder-level methods.

Method	RGB		DIN		Thermal		UV		Mean	
	pp	rel%	pp	rel%	pp	rel%	pp	rel%	pp	rel%
CMAG	56.61	66.99	51.13	61.30	35.78	42.78	29.37	34.81	43.22	51.47
GCMA	61.09	74.42	51.00	62.85	35.69	42.96	20.99	25.91	42.19	51.53
MWPA	67.22	81.01	50.44	61.02	24.54	29.71	12.32	14.81	38.63	46.64
PL-MMTM	58.76	73.55	52.51	65.69	11.31	14.11	4.72	5.87	31.82	39.81
PL-R2AU	55.43	66.48	47.26	57.13	19.06	23.18	7.35	8.90	32.27	38.92
PL-SIG	57.91	69.83	45.31	55.18	21.34	26.00	26.17	31.88	37.68	45.72
Decoder Mean	59.50	72.05	49.61	60.53	24.62	29.79	16.82	20.36	37.64	45.68
GF-Net	71.47	83.44	48.04	56.54	54.08	63.32	25.66	30.07	49.81	58.34

RGB removal causes severe degradation across both approaches (Table 10), yet decoder fusion retains substantially more performance than encoder fusion. Converting to residual performance, PL-SIG maintains 29.91% of its baseline mIoU under RGB drop (24.71 mIoU from an 82.62 baseline), nearly 2× higher than GF-Net's 16.43% retention (14.05 from an 85.52 baseline). This resilience advantage, 13.56 pp lower degradation for PL-SIG, demonstrates that modality isolation in decoder fusion enables more graceful degradation when the primary sensor fails. The decoder mean (59.50 pp RGB degradation) outperforms encoder fusion by 11.97 pp (71.47 pp - 59.50 pp), confirming that this architectural advantage extends across all decoder variants.
DIN removal reveals comparable vulnerabilities across paradigms (decoder mean: 49.61 pp vs encoder: 48.04 pp), with

Table 11: Performance comparison between decoder and encoder fusion under noise perturbations at intensity 2.5. Values show mIoU degradation (pp) for each modality-noise combination averaged across lighting conditions. The pivoted structure reveals modality-specific vulnerabilities: decoder fusion demonstrates superior RGB and thermal resilience, while encoder fusion's data-level DIN integration provides exceptional noise tolerance for depth features. GF-Net exhibits particularly severe thermal degradation (47.39 pp mean), exceeding decoder methods by 3.8×.

Method	Modality	Basic	Gaussian	Salt&Pepper	Speckle	Mean
CMAG (Decoder)	RGB	22.51	31.91	32.17	19.37	26.49
	DIN	3.43	6.60	8.54	5.24	5.95
	Thermal	11.09	10.45	14.87	14.12	12.63
	UV	16.10	19.04	19.51	11.90	16.64
PL-SIG (Decoder)	RGB	26.78	37.51	37.67	20.26	30.55
	DIN	2.08	3.90	4.68	3.05	3.43
	Thermal	9.85	9.48	12.82	12.43	11.14
	UV	6.52	8.42	8.03	4.28	6.81
GF-Net (Encoder)	RGB	38.57	50.33	47.93	34.35	42.80
	DIN	0.77	1.81	2.22	1.89	1.67
	Thermal	41.76	34.73	55.32	57.74	47.39
	UV	1.41	6.08	4.16	0.36	3.00

PL-SIG achieving the best decoder performance (45.31 pp). This minimal 1.57 pp difference reflects the fact that both approaches rely on SWIF-enhanced RGB+DIN features established at the encoder level; removing DIN degrades this shared foundation equally, regardless of where thermal and UV auxiliaries are subsequently integrated. The comparable impact confirms that depth-intensity-normals features are equally critical to both paradigms, as geometric cues are embedded early in the feature hierarchy before the fusion stage divergence.

Thermal and UV drops expose the most striking architectural differences. For thermal removal, encoder fusion suffers severe 54.08 pp degradation, more than double the decoder mean (24.62 pp) and 2.5× worse than PL-SIG (21.34 pp). This 29.46 pp gap reveals that early thermal integration creates brittle dependencies that cascade through the entire encoder when disrupted. Conversely, decoder methods exhibit wide variation in auxiliary resilience: PL-MMTM (11.31 pp) and PL-R2AU (19.06 pp) demonstrate superior thermal independence through channel gating and recurrent attention, while CMAG (35.78 pp) and GCMA (35.70 pp) show higher sensitivity due to explicit cross-modal dependencies.

Noise robustness comparisons at intensity 2.5 (Table 11) reveal the impact of fusion stage placement across four noise types. The overall mean degradation (unweighted average across modalities) shows decoder-level PL-SIG achieving 12.98 pp, substantially outperforming encoder-level GF-Net (23.71 pp) by 10.73 pp. RGB-targeted corruption induces the most severe degradation across all methods, with encoder-level GF-Net exhibiting the highest sensitivity (42.80 pp mean) compared to decoder-level CMAG (26.49 pp) and PL-SIG (30.55 pp). Notably, CMAG demonstrates superior RGB resilience through its cross-modal attention mechanism, which adaptively redistributes representational load when primary features degrade, outperforming PL-SIG by 4.06 pp.

Thermal corruption under noise exposure exhibits catastrophic failure in encoder fusion: GF-Net suffers a mean degradation of 47.39 pp, substantially exceeding decoder-level methods (CMAG: 12.63 pp, PL-SIG: 11.14 pp) by factors of 3.8× and 4.3×, respectively. This vulnerability is particularly pronounced under salt-and-pepper (55.32 pp) and speckle (57.74 pp) noise, demonstrating that early fusion of thermal features creates brittle dependencies that catastrophically fail under severe corruption. The extreme speckle degradation (57.74 pp) represents a near-total loss of thermal information, indicating that multiplicative noise fundamentally disrupts encoder-level feature interactions. The consistent thermal advantage of decoder fusion across all noise types confirms that late integration preserves modality independence to a degree, enabling a more graceful degradation when auxiliary sensors are compromised.

Conversely, DIN exhibits exceptional noise resilience across both paradigms due to the shared SWIF mechanism, though encoder fusion achieves marginally superior tolerance (GF-Net: 1.67 pp mean degradation vs PL-SIG: 3.43 pp, CMAG: 5.95 pp). Since both architectures identically fuse DIN with RGB at each encoder stage via SWIF, producing the same enhanced RGB+DIN primary stream, this small difference arises from downstream architectural choices rather than DIN processing itself. In encoder fusion, the SWIF-enhanced stream is immediately fused with thermal and UV at each encoder stage, allowing auxiliary modalities to interact with the robust RGB+DIN representation throughout the encoder depth. In decoder fusion, the SWIF-enhanced stream propagates independently through the encoder before

late fusion with auxiliaries, potentially accumulating slightly different noise characteristics. Similarly, UV corruption has minimal impact on GF-Net (3.00 pp mean), with speckle noise producing negligible degradation (0.36 pp), while decoder methods exhibit higher UV sensitivity (PL-SIG: 6.81 pp, CMAG: 16.64 pp). CMAG's attention-based fusion amplifies UV noise through explicit cross modal dependencies that propagate corruption across modalities at the decoder stage.

Across all decoder methods, mean degradation ranges from 14.3 pp to 23.0 pp, with conservative fusion strategies (PL-R2AU: 14.3 pp, PL-MMTM: 15.6 pp) achieving superior tolerance through residual connections and transfer modules that maintain independent gradient pathways. Attention-based methods (CMAG: 21.0 pp, MWPA: 23.0 pp) trade noise resilience for baseline accuracy, as their explicit cross-modal interactions create stronger dependencies that amplify corruption effects. Both paradigms degrade substantially under severe corruption, confirming that sensor noise mitigation remains an open challenge regardless of fusion strategy. However, the 10.73 pp decoder advantage demonstrates that architectural choices significantly impact robustness margins.

Figure 18 visualises class-specific resilience patterns, revealing that decoder-level PL-SIG maintains more consistent per-class IoU across perturbations compared to encoder-level GF-Net. The most pronounced differences emerge under modality drops (leftmost perturbation columns), where GF-Net exhibits systematic class collapse (extensive red regions) while PL-SIG preserves moderate discrimination (yellow-green regions). This visualisation confirms that late integration's modality independence translates to more uniform degradation across semantic categories, avoiding the catastrophic class-specific failures characteristic of early fusion when primary modalities fail.

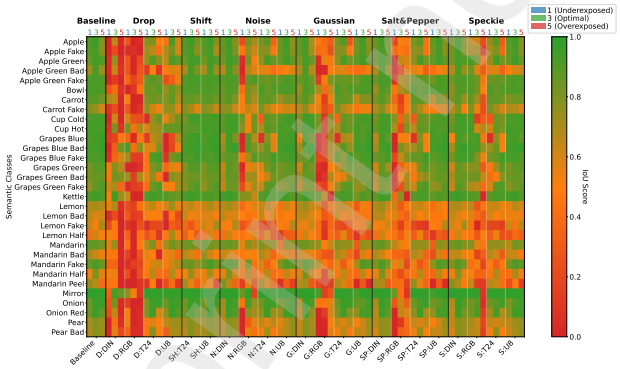
Architectural Trade-offs and Application Context. The comparative analysis establishes three deployment scenarios:

(i) **Controlled environments** with geometrically aligned sensors and strict latency constraints favour encoder fusion for peak accuracy (+3.23 pp vs PL-SIG) and efficiency (+62% throughput).

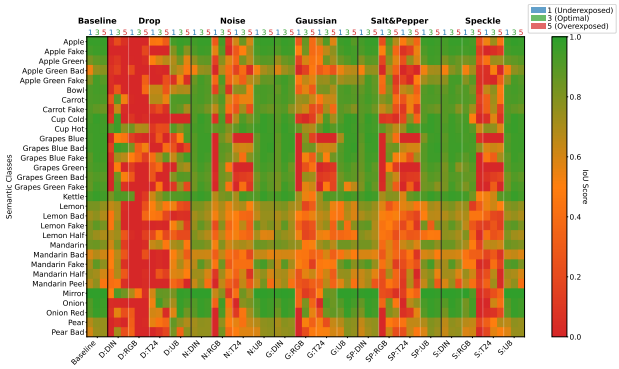
(ii) **High-accuracy scenarios** tolerating modest computational overhead benefit from CMAG's hybrid fusion, achieving competitive accuracy (within 2 pp) with inherent alignment tolerance.

(iii) **Robustness-critical applications** with potential sensor failures or geometric drift require decoder-level fusion, accepting 3 pp accuracy reduction for 2.0× improved sensor failure resilience (Section 6.7) and intrinsic spatial tolerance (Section 6.7.2).

The choice between paradigms depends on operational constraints: encoder fusion maximises performance under assured alignment and sensor reliability, while decoder fusion prioritises robustness for challenging conditions where failures and misalignment are anticipated.



(a) PL-SIG (decoder-level): Maintains class discrimination under drops/shifts through modality isolation.



(b) GF-Net (encoder-level): Shows systematic class collapse under RGB impairment, particularly for drops.

Figure 18: **Class-wise robustness comparison under perturbations.** Heatmaps show per-class IoU (green=high, red=low) across baseline and perturbation conditions (Drop, Shift, four noise types) under three RGB lighting conditions (1: underexposed, 3: optimal, 5: overexposed). Decoder-level PL-SIG preserves class-specific performance more consistently than encoder-level GF-Net, particularly under modality drops (leftmost perturbation group) where early fusion creates cascading failures. The systematic difference in drop columns demonstrates decoder fusion's architectural advantage in maintaining auxiliary pathway independence.

6.11. Discussion

6.11.1. Accuracy-Robustness Trade-Off

Architectural fusion strategies exhibit distinct performance-robustness profiles. CMAG achieves the highest mean accuracy (84.18% mIoU) but demonstrates elevated noise sensitivity (12.93 pp mean degradation), while MWPA exhibits the highest noise vulnerability (13.91 pp) despite mid-range baseline performance (82.89% mIoU). However, PL-R2AU demonstrates that aggressive accuracy-robustness trade-offs are not inevitable, achieving competitive accuracy (83.14% mIoU, only 1.04 pp below CMAG) while maintaining the lowest noise sensitivity (8.80 pp)—a 4.13 pp robustness advantage over CMAG. Similarly, PL-MMTM demonstrates balanced characteristics with the lowest baseline accuracy (79.89% mIoU) yet second-best noise robustness (9.67 pp), indicating that conservative fusion strategies can maintain robustness without substantial accuracy penalties. The divergent profiles stem from learnt fusion strategies: aggressive auxiliary exploitation (CMAG, MWPA) maximises discriminative capacity under ideal conditions but amplifies vulnerability when auxiliary channels degrade, whereas selective modality utilisation (PL-R2AU, PL-MMTM) maintains robustness through conservative fusion gains.

6.11.2. Modality Hierarchy and Architectural Invariance

Drop ablations establish a consistent modality importance hierarchy (RGB > DIN > T24 > U8) across all six architectures, with mean losses of 59.50 pp, 49.61 pp, 24.62 pp, and 16.82 pp, respectively, averaged across all architectures and lighting conditions. This invariance demonstrates that information content dominates architectural effects: RGB-DIN provides geometric structure and appearance, while thermal and UV contribute specialised discriminative cues for challenging classes (Section 6.5). However, fusion strategies substantially influence sensitivity magnitude: PL-MMTM achieves a 31.82 pp mean drop loss through modality isolation, while CMAG's cross-modal integration amplifies dependence (43.22 pp).

Thermal demonstrates greater importance than UV for fusion performance: when either auxiliary modality is removed, networks retain 57.8% mean residual mIoU without thermal compared to 65.6% without UV (averaged across all architectures and lighting conditions), a 7.8 percentage point difference that quantifies the relative contribution of each auxiliary modality to segmentation performance.

6.11.3. Decoder-Level Spatial Robustness

Spatial shift ablations reveal intrinsic misalignment tolerance: 20 px offsets cause only 2.87 pp mean degradation vs 37.64 pp for complete modality removal—a 13.1× difference. This resilience stems from decoder fusion's semantic-level integration, where spatial correspondence assumptions are relaxed compared to encoder-level pixel alignment. Thermal exhibits 1.63× greater shift sensitivity than UV, with vertical shifts proving most damaging.

6.11.4. Fusion Stage Selection and Operational Context

The encoder-decoder comparison (Section 6.10) reveals that the placement of the fusion stage represents a primary design decision. Encoder-level fusion maximises inference throughput (55 vs. 31–34 FPS) and computational efficiency (17.3 vs. 74.0–91.7 GFLOPs) through feature sharing, while decoder-level fusion prioritises robustness through modality isolation (retains 30.17% of baseline vs. 16.56% for encoder fusion under RGB loss). However, encoder-level fusion requires geometrically aligned inputs, imposing preprocessing overhead not reflected in the reported inference metrics. Thermal and UV images must undergo rectification and lens distortion correction. Whilst these operations execute in parallel for both modalities, conservative estimates based on typical performance for VGA-resolution thermal imagery suggest approximately 5 ms preprocessing latency per frame. Although the network maintains a 55 FPS inference capability, the mandatory preprocessing creates a 5 ms system latency and incurs additional computational costs (CPU-based rectification) before frames enter the GPU-accelerated network. In contrast, decoder-level fusion operates directly on raw, distorted sensor streams, processing frames immediately upon acquisition without the delay associated with registration preprocessing.

For controlled environments with mechanically stable sensor arrays and strict latency requirements, the accuracy advantage of encoder fusion (+1.34 pp mean across RGB1/3/5 vs. the best decoder method) may justify adoption if the 5 ms preprocessing latency and additional CPU overhead remain acceptable for the application. Conversely, for field robotics or scenarios with mechanical vibration, sensor degradation, or calibration drift, the more graceful degradation of decoder fusion and its intrinsic tolerance to misalignment outweighs the 1 to 2 percentage point accuracy

reduction. Additionally, decoder fusion eliminates the registration preprocessing requirement entirely, enabling zero-delay frame-to-prediction throughput, which is critical for reactive robotic control.

6.11.5. Limitations

Whilst CMAG demonstrates strong performance for unaligned multimodal fusion, several limitations warrant acknowledgement:

Computational overhead. CMAG carries a higher compute/parameter budget than the most efficient baseline (91.7 vs. 74.0 GFLOPs for PL-MMTM; ≈24% increase) and a larger parameter count (22M vs. 19M). Nevertheless, all methods sustain real-time inference (31–34 FPS). The attention path (GCMA) contributes most of this overhead.

Training complexity and memory. Using separate decoder heads per modality increases training-time memory and adds optimisation complexity. In particular, we employ per-head learning rates and multi-head supervision with loss weighting and residual warm-up. These settings improve stability and graceful degradation under missing inputs, but they also enlarge the hyperparameter search space and can make exact reproduction more sensitive to configuration.

Noise–accuracy trade-off. CMAG tends to yield larger absolute mIoU losses under severe corruptions than conservative fusion (e.g., PL-R2AU, PL-MMTM), while performing competitively at mild to moderate intensities. This reflects an explicit design choice; aggressive auxiliary utilisation maximises discriminative capacity in clean conditions but increases vulnerability when auxiliaries degrade (see Table 7).

Alignment tolerance bounds. Modality-level pooling in GCMA provides robustness to moderate misalignment; however, performance degrades progressively with spatial shifts. In our misregistration study, 20-pixel shifts cause modest degradation (2.61 pp average), while 40-pixel shifts induce more substantial losses (5.68 pp average; see Table 6). Performance under larger misalignments or non-translational distortions (e.g., rotation, scale) remains untested and warrants further investigation.

Domain scope. Results are reported on MM5 (indoor produce with controlled RGB lighting and auxiliary thermal/UV). Generalisation to other domains — e.g., outdoor scenes, autonomous driving, or medical imaging with different sensor suites — remains to be established.

Calibration and failure awareness. The present model does not include explicit confidence calibration or lightweight sensor-health cheques (e.g., dropout, drift detection). Integrating uncertainty quantification and simple failure detectors would better support safety-critical or time-critical deployments.

6.11.6. Future Work

Several promising directions emerge from this study. The observed accuracy-robustness trade-off motivates training strategies aimed at shifting the optimisation frontier, including corruption-aware objectives, adversarial perturbation schemes targeted at auxiliary streams (thermal/UV), and multi-objective searches that balance clean accuracy against robustness to noise, modality drop, and spatial shift. Targeted augmentation—especially spatial perturbations reflecting realistic misalignment and modality-specific noise processes—may further enhance resilience while preserving clean-data performance.

The modality-specific contributions observed suggest potential for class-adaptive fusion that selectively weights inputs by semantic context at the pre-logit stage. Extending evaluation to other multimodal domains (e.g., autonomous driving, medical imaging) would test whether the design principles identified here generalise beyond controlled inspection settings.

Computational efficiency could be improved through lightweight attention mechanisms or knowledge distillation, addressing the 24% overhead while maintaining accuracy for edge deployment. Incorporating differentiable spatial transformation networks within the decoder could extend the alignment tolerance without sacrificing the benefits of late fusion.

Finally, integrating uncertainty quantification for calibrated confidence and lightweight failure detection for sensor malfunctions (e.g., auxiliary dropout or drift) would support safer deployment in time-critical applications, particularly when operating on unaligned and optically uncorrected auxiliary streams.

7. Conclusion

This work presents a comprehensive investigation of decoder-level fusion strategies for multimodal semantic segmentation using unaligned RGB+DIN, thermal, and UV imagery. We propose CMAG (Cross-Modal Attention

with Gated Residuals). This decoder-level fusion module combines global cross-modal attention with sigmoid-gated residuals to enable alignment-tolerant fusion without explicit geometric calibration. Through systematic evaluation of CMAG against five adapted baseline methods across three lighting conditions, we establish performance baselines, quantify modality contributions via ablation studies, and assess robustness to sensor noise across 4,032 configurations (four corruption types, fourteen intensity levels). Crucially, thermal and UV modalities are fused in their distorted form without lens correction, testing decoder-level fusion's capacity to handle realistic sensor imperfections alongside spatial misalignment.

Our findings reveal critical insights for decoder-level multimodal fusion design. Our proposed CMAG achieves the highest baseline accuracy (84.18% mIoU average, 87.61% under optimal lighting) through hybrid channel-modality attention gating at the decoder level, while maintaining moderate noise tolerance (12.93 pp mean degradation across all noise scenarios) and graceful lighting adaptation. GCMA (CMAG's attention component evaluated standalone) achieves strong optimal-lighting performance (86.72%) via cross-modal attention but exhibits increased sensitivity to suboptimal illumination (RGB1: 80.49%, RGB5: 78.03%). The baseline architectures demonstrate alternative trade-offs: PL-MMTM (adapted Multimodal Transfer Module) attains the best ablation robustness (31.82 pp mean drop loss, 1.07 pp shift loss) via squeeze-and-excitation fusion; PL-R2AU (adapted Recurrent Residual Attention U-Net) achieves the best noise resilience (8.80 pp mean loss, rising from 6.51 pp at mild corruption to 16.74 pp at severe levels) through recurrent attention mechanisms.

Ablation studies confirm a clear modality hierarchy (RGB > DIN > T24 > U8), with RGB removal causing a 59.50 pp average degradation (72.05% relative), and DIN showing critical importance under challenging lighting conditions (up to 63.99 pp loss when removed under underexposure). Thermal and UV modalities provide specialised discriminative information essential for challenging classes (fake objects, partially decayed fruit) despite modest overall importance (24.62 pp and 16.82 pp drop impacts, respectively). Notably, decoder-level fusion of unaligned modalities demonstrates strong spatial robustness, with 20 px misalignment causing only 2.11 pp average degradation compared to 37.64 pp for complete modality removal—indicating that moderate calibration drift poses minimal risk to segmentation accuracy at the decoder level.

Noise robustness evaluation reveals that RGB-targeted corruption induces the most severe degradation (Gaussian RGB: 31.91%–37.51% impact across networks; salt-and-pepper RGB: 29.31%–38.22%), while DIN demonstrates surprising noise tolerance (3.90%–7.55% average impact) despite being the second-most-important modality. Architectural differences in noise handling amplify under severe corruption: at mild intensity (i=1.0), network losses span a modest 2.98 pp range (6.51–9.49 pp), expanding to 10.30 pp at severe levels (i=5.0: 16.74–27.04 pp)—a 3.5× amplification.

Networks with higher baseline accuracy demonstrate greater noise sensitivity, while robust architectures sacrifice peak accuracy, indicating that architectural designs must balance clean-data performance against perturbation resilience. Comparison with encoder-level fusion (GF-Net) demonstrates that this trade-off extends across fusion paradigms: encoder integration achieves higher baseline accuracy and superior efficiency (55 FPS, 17.3 GFLOPs) through early feature sharing, while decoder-level designs (31–34 FPS, 74.0–91.7 GFLOPs) demonstrate stronger resilience under sensor failure (e.g., CMAG retains 32.75% vs. 16.43% of baseline under RGB drop). However, encoder-level fusion requires per-frame geometric alignment of auxiliary modalities, typically incurring approximately 5 ms CPU preprocessing latency at VGA resolution. Decoder-level fusion eliminates this preprocessing requirement by operating directly on raw, geometrically uncorrected sensor streams, reducing system latency and simplifying deployment. Architecture selection follows naturally from these trade-offs: encoder-level fusion (GF-Net) for controlled, high-throughput scenarios with reliable sensors where preprocessing overhead is acceptable; decoder-level fusion (CMAG, GCMA) when prioritising accuracy under modest perturbation; and robust decoder variants (PL-R2AU, PL-MMTM) when robustness is critical, with potential sensor degradation or misalignment. All decoder-level architectures achieve real-time inference (31–34 FPS), enabling practical robotic vision applications without requiring pre-aligned or geometrically corrected sensor inputs.

This work establishes comprehensive benchmarks for decoder-level multimodal fusion with unaligned, optically uncorrected inputs, providing empirical guidance for architecture selection and highlighting fundamental trade-offs in decoder-level fusion design. We acknowledge that our findings are validated exclusively on the MM5 dataset, comprising indoor produce inspection under controlled lighting variations. Generalisation of our learnt weighting patterns to outdoor environments, medical imaging, or autonomous driving with different sensor combinations remains to be explored empirically.

List of Abbreviations

ATT	Channel and spatial dual attention
BAM	Block Attention Module
CANet	Co-Attention Network
CBAM	Convolutional Block Attention Module
CMAF	Cross-Modal Attention Fusion
CMAG	Cross-Modal Attention with Gated Residuals
CMNeXt	Cross-Modal Next
CMX	Cross-Modal X
CNN	Convolutional Neural Network
CPS	Cross-Modal Prototype Sharing modules (in TCPSNet context)
CT	Computed Tomography
D	Depth
DGFM	Dual Gate Fusion Module
DGFNet	Dual Gate Fusion Network
DIN	Depth, Intensity, and Normals (MM5 Dataset)
DSM	Digital Surface Model
ETFormer	Edge-Thermal Transformer
FCN	Fully Convolutional Network
FEM	Feature Enhancement Module
FLAIR	Fluid-Attenuated Inversion Recovery
FPS	Frames Per Second
FRM	Feature Rectification Module
FWIoU	Frequency Weighted Intersection over Union
GAP	Global Average Pooling
GCMA	Global Context Modality Attention
GF-Net	Gated Fusion Network
GMFNet	Gated Multimodal Fusion Network
GN	Group Normalisation
GT	Ground Truth
HSI	Hyperspectral Imaging
HRNet	High-Resolution Network
LF-DLM	Late Fusion Deep Learning Model
LiDAR	Light Detection and Ranging
LN	Layer Normalisation
LWIR	Long-Wave Infrared
MCAM	Multi-scale Cross Attention Module
MEFNet	Modality Expert Fusion Network
MGFNet	Multi-Gated Fusion Network
MiT	Mix Transformer
mIoU	mean Intersection over Union
MLP	Multi-Layer Perceptron
MMTM	Multimodal Transfer Module
MPA	Mean Pixel Accuracy
MRI	Magnetic Resonance Imaging
MUUFL	Multi-sensor Urban/Unstructured Fusion and Learning
MWPA	Modality-wise Parallel Attention
NIR	Near-Infrared
PA	Pixel Accuracy
PET	Positron Emission Tomography
PICNet	Prototype-based Incremental Classification Network
PL	Pre-Logit
PL-ATT	Pre-Logit Channel and Spatial Dual Attention
PL-MMTM	Pre-Logit Multimodal Transfer Module
PL-R2AU	Pre-Logit Recurrent Residual Attention U-Net
PL-SIG	Pre-Logit Sigmoid Gating
PR	Primary (RGB+DIN stream)
PSPNet	Pyramid Scene Parsing Network
QSF-Net	Quality-aware Selective Fusion Network
R2AU	Recurrent Residual Attention U-Net
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
RGB-D	RGB-Depth
RGB-T	RGB-Thermal
SAR	Synthetic Aperture Radar
SE	Squeeze-and-Excitation
SGFNet	Semantic Guidance Fusion Network
SIG	Sigmoid-Gated (residuals)
SSMA	Self-Supervised Model Adaptation
SWIF	Stage-Wise Intensity Fusion
T	Thermal
T1	T1-weighted (MRI)
T1ce	T1-weighted contrast-enhanced (MRI)
T2	T2-weighted (MRI)
T24	Thermal 24-bit (MM5 Dataset)
TCPSNet	Two-stage Cross-modal Prototype Sharing Network
TH	Thermal (auxiliary stream)
U8	Ultraviolet 8-bit (MM5 Dataset)
UCTNet	Uncertainty-aware Cross-modal Transformer Network
UDFNet	Uncertainty-aware Dynamic Fusion Network
UV	Ultraviolet
VHR	Very High Resolution

Code Availability

The code used in this paper will be made publicly available at `https://github.com/martinbrennernz/MM5-Dataset` upon publication of this work.

Appendix A. Implementation Details

Appendix A.1. Normalisation details

Definitions.. Given pyramid features $P \in \mathbb{R}^{B\times C\times H\times W}$, LayerNorm (LN) normalises per instance over all channels and spatial positions,

$$\mathrm{LN}(P) = \gamma \odot \frac{P - \mu_{\mathrm{LN}}}{\sqrt{\sigma^2_{\mathrm{LN}} + \epsilon}} + \beta, \quad \mu_{\mathrm{LN}} = \frac{1}{CHW} \sum_{c,h,w} P_{b,c,h,w}.$$

GroupNorm (GN) partitions channels into G groups and normalises within each group:

$$\mathrm{GN}(P) = \gamma \odot \frac{P - \mu_{\mathrm{GN}}}{\sqrt{\sigma^2_{\mathrm{GN}} + \epsilon}} + \beta, \quad \mu^{(g)}_{\mathrm{GN}} = \frac{1}{(C/G)HW} \sum_{c\in\mathcal{G}_g,h,w} P_{b,c,h,w}.$$

Complexity remarks.. Both LN and GN have $O(BCHW)$ work per instance; GN exposes more parallelism by reducing over groups of size C/G. For C=512, H=480, W=640, LN reduces over 157.3×10^6 elements, while GN-16 reduces over 9.83×10^6 elements per group in parallel. Empirically, after training with LN, replacing LN by GN-16 at evaluation yields small relative deviations, $\|\mathrm{LN}(P) - \mathrm{GN}_{16}(P)\|_2/\|P\|_2 \approx 10^{-2}$, while improving throughput (Section 5).

Appendix B. Detailed Network Results

Appendix B.1. Class level results at 220 epochs

Table B.12: Detailed class-level network comparison across lighting conditions and fusion architectures. Shows per-class IoU, overall metrics (mIoU, FIoU, MPA, PA), and mean rank scores. Best value per RGB configuration highlighted in bold. RGB1: underexposed; RGB3: optimal; RGB5: overexposed lighting. "Bad" classes are partially rotten; "Fake" classes are replicas.

Class	RGB1 CMAG	RGB3 CMAG	RGB5 CMAG	RGB1 GCMA	RGB3 GCMA	RGB5 GCMA	RGB1 PL-R2AU	RGB3 PL-R2AU	RGB5 PL-R2AU	RGB1 MWPA	RGB3 MWPA	RGB5 MWPA	RGB1 PL-SIG	RGB3 PL-SIG	RGB5 PL-SIG	RGB1 PL-MMTM	RGB3 PL-MMTM	RGB5 PL-MMTM
Apple	91.37	**96.72**	69.22	90.58	96.26	76.96	**92.81**	96.09	87.56	84.80	95.32	**90.06**	90.30	96.02	78.54	91.92	96.11	88.95
Apple Fake	89.23	93.81	78.30	89.52	93.36	82.75	**91.48**	**94.34**	88.07	85.13	92.97	**88.60**	88.78	93.64	76.54	90.22	93.46	88.55
Apple Green	82.79	**95.80**	94.01	**85.91**	94.38	84.31	80.06	94.48	93.08	82.39	95.55	**94.70**	85.68	94.52	90.72	78.34	93.55	83.31
Apple Green Bad	63.07	92.38	89.28	71.97	91.49	63.75	60.01	**94.26**	75.39	64.24	90.60	**91.52**	**72.17**	92.93	73.07	60.84	92.11	71.50
Apple Green Fake	90.58	91.90	**94.17**	**93.08**	**92.77**	91.13	91.37	92.73	85.87	90.65	90.64	91.94	90.97	92.24	91.58	92.78	90.85	89.81
Background	**99.81**	**99.86**	**99.81**	99.79	**99.86**	99.81	99.79	99.85	99.80	99.78	99.84	99.80	99.80	99.84	99.79	99.77	99.84	99.77
Bowl	**92.29**	**93.22**	91.13	90.90	92.53	**92.48**	91.87	92.13	90.03	90.01	91.97	90.42	91.93	92.15	91.80	90.23	91.48	89.35
Carrot	87.86	**90.05**	87.05	85.70	84.89	86.30	86.57	86.14	86.80	86.43	88.04	86.90	**88.23**	88.20	83.53	83.19	86.77	**87.41**
Carrot Fake	75.64	**79.74**	72.15	72.42	59.38	72.42	72.13	65.52	71.27	76.51	76.11	69.97	**80.90**	74.56	60.66	64.97	68.56	**75.70**
Cup Cold	90.00	94.98	93.58	91.74	94.66	90.94	**92.12**	94.60	**95.08**	82.90	**95.06**	94.11	91.66	93.96	94.07	91.89	92.47	93.91
Cup Hot	93.67	**94.75**	93.23	93.41	94.06	91.54	**94.46**	94.29	93.72	88.62	93.08	**93.88**	94.31	93.83	93.26	91.12	92.59	92.82
Grapes Blue	**94.24**	94.38	**95.92**	84.60	95.77	72.14	93.41	95.96	95.46	90.55	**96.10**	95.17	91.34	94.72	95.35	90.04	93.95	89.95
Grapes Blue Bad	93.71	**96.14**	96.25	93.70	95.88	96.63	93.66	93.64	**96.66**	**94.03**	93.80	96.15	93.28	94.76	96.24	93.39	93.63	94.85
Grapes Blue Fake	91.52	95.62	**95.38**	89.13	**96.24**	84.05	79.40	95.50	94.89	91.68	95.20	95.32	**92.42**	91.84	94.58	79.56	95.04	91.73
Grapes Green	81.85	89.08	**90.99**	72.32	88.19	89.27	79.80	86.52	85.76	**87.41**	88.16	90.13	81.04	**89.54**	89.84	85.87	87.86	86.60
Grapes Green Bad	**84.22**	**87.69**	84.80	81.03	87.21	84.99	81.31	83.21	85.10	82.18	86.36	**86.22**	81.55	85.57	85.45	81.94	86.24	83.81
Grapes Green Fake	82.27	**92.75**	90.54	74.74	91.19	81.49	74.11	91.83	88.63	**82.95**	92.04	85.92	75.64	87.40	**90.87**	80.51	91.36	86.41
Kettle	**92.44**	**95.87**	94.24	91.01	95.66	**95.08**	90.82	95.77	93.77	91.36	94.49	94.32	91.86	94.77	91.74	89.28	92.61	94.54
Lemon	**70.98**	**77.91**	**68.97**	66.69	75.35	66.26	66.15	72.71	66.46	66.95	72.24	64.33	68.66	71.13	65.58	69.00	69.27	63.94
Lemon Bad	**68.55**	72.04	60.28	58.89	**72.29**	54.00	60.75	67.22	57.02	60.28	64.47	53.08	66.70	58.54	**60.92**	64.80	60.94	49.74
Lemon Fake	63.60	**70.82**	**70.72**	**74.34**	60.85	58.98	59.35	63.73	61.17	56.02	62.48	58.97	63.71	52.50	70.19	68.56	41.13	54.13
Lemon Half	54.65	**67.21**	59.96	47.12	64.78	54.22	58.70	63.55	**60.70**	49.37	66.47	56.81	41.10	61.68	49.93	**60.48**	59.89	54.46
Mandarin	85.71	84.09	82.24	85.23	87.20	75.88	86.18	86.39	78.22	85.71	**87.24**	80.38	86.22	86.20	**82.32**	**87.18**	85.29	81.36
Mandarin Bad	66.98	54.36	53.85	61.04	68.31	23.14	70.22	**77.74**	38.27	69.10	64.72	43.65	72.62	62.99	**54.37**	**73.22**	58.57	45.82
Mandarin Fake	85.93	92.38	74.98	84.19	91.63	80.98	79.94	86.38	80.58	**87.29**	**92.72**	**84.73**	85.25	92.36	84.55	85.67	91.32	83.94
Mandarin Half	62.86	82.20	68.26	60.95	**83.15**	71.62	**85.48**	80.79	**73.33**	44.89	71.57	55.93	52.07	80.77	56.75	71.15	78.08	54.56
Mandarin Peel	81.90	78.74	56.35	**84.36**	**79.80**	54.39	80.16	64.76	**64.31**	53.34	60.55	33.50	75.97	66.13	38.28	63.40	66.06	31.70
Mirror	**98.75**	98.89	98.64	98.64	**98.96**	**98.68**	98.62	98.91	98.37	98.40	98.87	98.52	98.64	98.67	98.57	98.20	98.86	98.19
Onion	**94.34**	**96.63**	**95.22**	83.62	96.35	95.04	84.39	96.19	95.03	82.34	94.43	94.92	83.96	96.19	95.11	85.18	95.75	94.23
Onion Red	**93.31**	**96.18**	84.56	83.65	95.66	84.72	83.23	95.72	**95.03**	83.00	94.16	94.20	83.89	95.63	86.62	82.04	95.35	92.51
Pear	69.63	**79.01**	76.66	69.50	78.78	72.01	**73.20**	76.00	75.14	71.23	78.76	**77.07**	71.63	78.44	74.63	71.21	75.94	74.31
Pear Bad	67.43	78.19	75.56	65.77	78.17	71.06	**72.79**	75.56	73.97	68.19	**78.73**	**75.96**	68.54	78.29	73.99	68.28	75.93	74.55
Mean Rank	**10.3**	**3.1**	**9**	12.6	4.6	12.1	11.8	5.6	10.5	13.5	5.9	9.9	10.9	6.2	10.9	12.6	8.5	13.1
Overall mIoU	**82.54**	**87.61**	**82.38**	80.49	86.72	78.03	81.39	86.02	82.02	78.99	85.71	81.47	81.27	85.31	80.29	80.76	84.09	79.45
Overall FIoU	**99.37**	**99.55**	**99.37**	99.29	99.53	99.28	99.30	99.51	99.35	99.26	99.50	99.36	99.33	99.48	99.32	99.27	99.46	99.28
Overall MPA	**89.21**	**92.58**	**88.98**	88.00	91.87	85.18	88.13	91.86	88.21	86.21	91.70	87.30	88.75	91.45	87.04	87.76	90.17	86.54
Overall PA	**99.66**	**99.76**	**99.66**	99.61	99.75	99.61	99.62	99.74	99.65	99.60	99.73	99.65	99.64	99.72	99.63	99.61	99.70	99.61
FPS	31	31	31	31	31	31	34	34	34	32	32	32	34	34	34	34	34	34
Parameters	22M	22M	22M	22M	22M	22M	19M	19M	19M	19M	19M	19M	19M	19M	19M	19M	19M	19M
GFLOPs	91.7	91.7	91.7	89.2	89.2	89.2	79.0	79.0	79.0	84.0	84.0	84.0	84.0	84.0	84.0	74.0	74.0	74.0

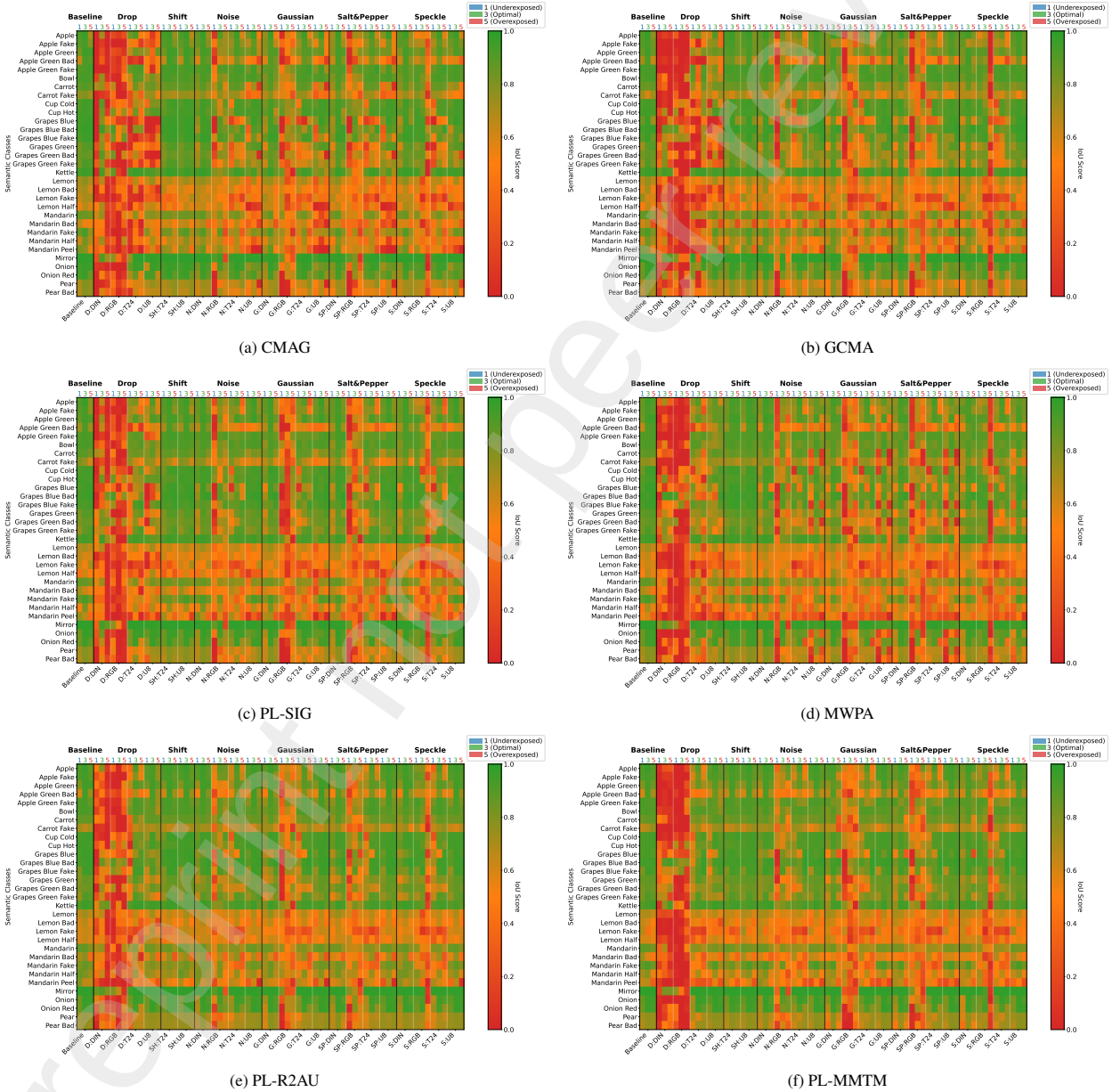
Appendix B.2. Network drop ablation results

Table B.13: Drop ablation results (detailed): residual mIoU (%) after removing each modality, split by lighting condition (RGB1/3/5). Networks ordered by robustness (left to right: least to most robust). Lower residual values indicate greater modality dependence.

Dropped	**RGB**	**CMAG**	**GCMA**	**MWPA**	**PL-SIG**	**PL-R2AU**	**PL-MMTM**
RGB	RGB1	34.51	30.09	11.81	37.19	31.09	19.75
	RGB3	17.49	11.73	11.81	11.91	19.41	21.80
	RGB5	30.70	20.15	23.38	25.04	32.65	21.80
DIN	RGB1	18.54	24.01	27.63	27.89	33.40	22.05
	RGB3	57.86	45.37	42.10	56.16	52.26	32.17
	RGB5	22.73	22.86	27.63	27.89	21.98	27.91
T24	RGB1	36.30	47.97	62.48	55.85	48.07	63.23
	RGB3	61.99	31.51	65.74	72.15	77.61	74.60
	RGB5	46.89	58.68	46.83	55.85	66.57	67.92
U8	RGB1	55.48	56.48	70.57	50.03	74.44	70.61
	RGB3	57.13	54.58	70.57	60.38	77.52	78.88
	RGB5	51.93	71.24	70.57	58.98	75.72	75.17
Mean loss (pp)		**43.22**	**42.19**	**38.63**	**37.68**	**32.27**	**31.82**

Appendix B.3. Decoder-Level Robustness Comparison

Figure B.19 presents comprehensive class-wise robustness heatmaps for all six decoder-level fusion architectures evaluated in this work. Each heatmap visualises per-class IoU (green=high, red=low) across baseline and perturbation conditions (Drop, Shift, four noise types) under three RGB lighting conditions (1: underexposed, 3: optimal, 5: overexposed). The layouts compare architecturally related methods: attention-based mechanisms (CMAG vs GCMA), lightweight gating versus parallel attention (PL-SIG vs MWPA), and conservative fusion strategies (PL-R2AU vs PL-MMTM). Conservative methods (bottom row) exhibit more uniform performance across perturbations, while attention-based approaches (top row) achieve higher baseline performance with increased vulnerability under severe drops. All decoder methods demonstrate superior modality isolation compared to encoder-level fusion (Figure 18 in the main text), with drop columns showing substantially less systematic class collapse.



(a) CMAG

(b) GCMA

(c) PL-SIG

(d) MWPA

(e) PL-R2AU

(f) PL-MMTM

Figure B.19: Class-wise robustness heatmaps for decoder-level fusion architectures under perturbations at intensity 2.5.