LEARNING TASK-AGNOSTIC REPRESENTATION FOR MULTIMODAL IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent image fusion methods increasingly incorporate downstream tasks into joint optimisation, with the aim of injecting task-oriented semantic information into the fused image. However, compressing multimodal information into a single fused image inevitably imposes a fixed trade-off between the source modalities, making it difficult to accommodate the diverse information requirements of different scenes and downstream tasks. In this work, we explore an alternative paradigm. Rather than compressing complementary multimodal cues into a single fused representation, we aggregate source information under complementary fusion objectives to learn *Task-Agnostic Representations* (TAR). By allocating heterogeneous fusion cues that would otherwise compete within a shared output space, TAR preserves richer complementary information and provides a flexible intermediate representation for diverse downstream applications, while maintaining the same information budget as conventional image-fusion outputs. Importantly, TAR is learned without relying on task-specific annotations or downstream-task supervision, allowing the same representation to be directly exploited by different perception tasks. For conventional image fusion, a lightweight fusion head can further project TAR into a standard fused image, providing a flexible means of recovering visually faithful fusion results. Extensive experiments across multiple fusion and perception benchmarks demonstrate that TAR consistently improves downstream object detection and semantic segmentation performance while retaining competitive image-level fusion quality.

1 INTRODUCTION

Multimodal image fusion aims to integrate complementary information from different sensing modalities into a unified representation. For infrared and visible image fusion, this task endeavors to preserve illumination-robust thermal saliency from infrared images and salient texture details from visible images. Beyond human visual enhancement, fused images are increasingly used as intermediate representations for downstream perception systems, including object detection and semantic segmentation Ma et al. (2019); Li & Wu (2018), with applications in nighttime autonomous driving, remote sensing, and video surveillance Li et al. (2022); Jin et al. (2017); Cheng et al. (2025).

Image fusion methods have made substantial progress, but most share a common problem formulation: complementary information from different modalities is compressed into a single fused image. Traditional methods rely on hand-crafted priors and predefined fusion rules Toet (1989); Rockinger (1997); Zhang et al. (2018); Li & Wu (2017), while recent learning-based methods using CNNs, Transformers, and diffusion models have further improved fusion quality Zhang et al. (2020); Ma et al. (2022); Zhao et al. (2023b); Zhang et al. (2026). However, both paradigms typically optimise a single fused output with multiple image-level objectives, which imposes a fixed trade-off between source modalities and makes it difficult to satisfy the diverse information requirements of different scenes and downstream tasks.

This limitation becomes more pronounced when fused images are fed to downstream perception networks. A single fused image must simultaneously encode infrared thermal responses, visible edge details, and cross-modal structural consistency, although these cues may favour different image characteristics. Moreover, because common detection and segmentation networks adopt RGB-format inputs, single-channel fused images are often replicated or simply mapped into three channels, pro-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

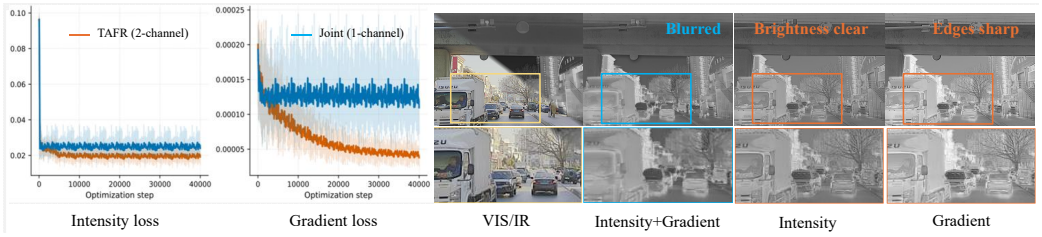

Figure 1: As shown on the left, our task-specialized formulation (orange) converges faster and to a lower loss value than the entangled single-channel formulation (blue), since each channel is optimised with its dedicated objective without being interfered by other objectives. The right side further shows that the specialized channels encode more semantically meaningful information with higher density: the intensity channel preserves clearer brightness distribution of thermal targets, while the gradient channel retains sharper edge details. In contrast, the joint intensity+gradient output must compromise between multiple objectives and appears blurred and less informative, confirming the advantage of task-agnostic representation.

ducing pseudo-RGB representations with high channel redundancy and no explicit functional roles. As a result, high image-level fusion quality does not necessarily translate into better downstream perception; natural-looking outputs may suppress thermal saliency, object boundaries, or semantic structures that are critical for machine perception Zhang et al. (2025).

More formally, given a single fused output y, existing image-level fusion objectives are usually imposed on the same representation:

$$\mathcal{L}_{\text{single}} = \mathcal{L}_{\text{I}}(y) + \mathcal{L}_{\text{G}}(y) + \mathcal{L}_{\text{S}}(y), \tag{1}$$

where $\mathcal{L}_{\text{I}}$, $\mathcal{L}_{\text{G}}$, and $\mathcal{L}_{\text{S}}$ denote intensity, gradient, and structural objectives, respectively. Such a shared-output formulation forces heterogeneous fusion cues to compete within the same output space, because three objectives with different preferred image characteristics must be satisfied by one output channel. As shown in Fig. 1, under matched training conditions, the single-channel formulation exhibits less favorable convergence for intensity and gradient losses than the specialized formulation, precisely because these heterogeneous objectives compete for the same output channel. The representative outputs further show that the single-channel result provides a compromised representation, whereas specialized intensity and gradient channels preserve complementary brightness and edge information more explicitly. For clarity, Fig. 1 visualizes the intensity and gradient channels, while the structure channel is included in the complete representation.

An alternative line of research introduces downstream supervision into the fusion process to improve perception performance. While effective for a specific task, such approaches typically couple the fusion model with task-specific annotations or optimisation objectives, limiting cross-task generality. In contrast, our approach learns the fusion representation without any downstream supervision.

To address these limitations, we propose Task-Agnostic Representations (TAR), a task-agnostic representation learning paradigm for multimodal image fusion, and instantiate it on infrared–visible fusion. Rather than compressing complementary multimodal cues into a single fused representation, TAR aggregates source information under complementary fusion objectives and assigns heterogeneous cues to role-specific channels. In our implementation, the three channels are associated with intensity, gradient, and structure roles, respectively:

$$\mathcal{L}_{\text{specialized}} = \mathcal{L}_{\text{I}}(Z_{\text{I}}) + \mathcal{L}_{\text{G}}(Z_{\text{G}}) + \mathcal{L}_{\text{S}}(Z_{\text{S}}), \tag{2}$$

where Z_{I}, Z_{G}, and Z_{S} denote the intensity-, gradient-, and structure-aware channels. By allocating heterogeneous fusion cues that would otherwise compete within a shared output space, TAR preserves richer complementary information and provides a flexible intermediate representation for diverse downstream applications, while maintaining the same three-channel information budget as conventional RGB-format inputs. Importantly, TAR is learned without task-specific annotations, downstream-task supervision, or perception heads, allowing the same representation to be directly exploited by different perception tasks. Since TAR only changes the output representation and its corresponding optimisation objectives, it can be plugged into different fusion backbones, including CNN- and Transformer-based networks, by replacing the conventional single-image output head

with a role-specific representation head. For conventional image fusion, a lightweight fusion head further projects TAR into a standard fused image, providing a flexible means of recovering visually faithful fusion results. Extensive experiments across multiple fusion and perception benchmarks demonstrate that TAR improves downstream object detection and semantic segmentation performance while retaining competitive image-level fusion quality.

The main contributions of this paper are as follows:

- We propose Task-Agnostic Representations (TAR), a role-specific three-channel representation for multimodal image fusion. By assigning intensity, gradient, and structural objectives to separate output channels, TAR reduces the competition among heterogeneous fusion cues and preserves richer complementary information.
- We introduce a task-agnostic training paradigm that learns the fusion representation without downstream task annotations or task-specific fusion losses. The resulting output preserves the standard three-channel interface and can be directly used to retrain common detection and segmentation networks.
- We show that the proposed task-agnostic representation is architecture-agnostic and can be integrated into different fusion backbones, including CNN- and Transformer-based networks. Retraining these networks under the proposed paradigm leads to improved downstream detection and segmentation performance in our experiments.
- We demonstrate the effectiveness of TAR on three fusion benchmarks, including LLVIP, MSRS, and TNO, and two downstream perception benchmarks, including M^3FD for object detection and FMB for semantic segmentation. The experiments show that the proposed representation improves downstream perception while maintaining competitive image-level fusion quality.

2 Related Work

2.1 Representation Learning for Infrared–Visible Image Fusion

The representation of information in infrared–visible image fusion has evolved from handcrafted component decomposition to deep feature learning. Early methods mainly relied on signal-processing priors to construct component-wise or subspace representations. For example, Laplacian pyramids Toet (1989), wavelet transforms Rockinger (1997), and nonsubsampled contourlet transforms Da Cunha et al. (2006) decompose source images into complementary components at different scales or frequency bands. Sparse-representation methods Zhang et al. (2018) and low-rank-representation methods Li & Wu (2017) instead exploit sparsity and low-rank structures to model the source images. Although these methods can separate image components to some extent, their fusion processes typically depend on handcrafted rules. Moreover, such component decomposition does not necessarily correspond to semantic allocation with explicit functional roles. As a result, these methods have difficulty adapting to the heterogeneous characteristics of infrared and visible modalities and often exhibit limited generalisation under complex scenes and varying imaging conditions.

Recent learning-based methods employ CNNs, Transformers, and diffusion models to perform end-to-end feature extraction and fusion Zhang et al. (2020); Ma et al. (2022); Zhao et al. (2023b); Zhang et al. (2026). Compared with handcrafted approaches, these methods can learn more flexible cross-modal interactions directly from data and have achieved substantial improvements in fusion quality. However, most existing methods learn latent features through reconstruction, perceptual, or image-level fusion objectives, without explicitly constraining the semantic roles of the output channels. Consequently, different channels may encode highly redundant information or uncontrollably mix complementary cues such as intensity, edges, and structural content. Although such representations can generate visually pleasing fused images, they are not necessarily suitable as direct inputs for downstream perception models.

In contrast, we formulate the fusion output as a three-channel task-agnostic representation with explicit channel semantics. Specifically, the three channels are constrained to encode intensity, gradient, and structural information, respectively, so that salient, detailed, and structural cues complementary between infrared and visible images can be preserved explicitly. This design avoids

compressing all multimodal information into a single fused image or redundant pseudo-RGB channels and provides standard downstream networks with a more structured three-channel input.

2.2 Image Fusion for Downstream Perception

Conventional image fusion methods primarily aim to improve the visual quality of fused images and are commonly evaluated using image-level metrics, such as SSIM, CC, PSNR, EN, and SCD Ma et al. (2019); Li & Wu (2018). This evaluation paradigm is appropriate when the fused result is mainly intended for human visualization. However, image-level quality does not always correlate with machine perception performance. Recent studies have shown that visually more natural or sharper fused images do not necessarily yield better downstream detection and segmentation results. Optimising excessively for visual appearance may even suppress thermal saliency, object boundaries, and semantic structures that are more important for machine perception Zhang et al. (2025).

To narrow the gap between image fusion and downstream perception, recent task-aware fusion methods have introduced task information into the fusion process. Existing approaches can be broadly divided into two categories. The first category designs task-related loss functions and propagates detection or segmentation supervision back to the fusion network, encouraging the fused image to better serve a particular downstream objective Tang et al. (2022); Liu et al. (2022); Zhao et al. (2023a); Bai et al. (2025); Liu et al. (2025). The second category directly injects task features from downstream networks into the fusion process to enhance the semantic expressiveness of the fused image Sun et al. (2022); Zhang et al. (2024); Zhao et al. (2025); Yang et al. (2026). These methods demonstrate the effectiveness of downstream supervision for improving task performance, but they also tightly couple the fusion model with task-specific annotations, task heads, and optimisation objectives. When the downstream objective changes from detection to segmentation, or when one fusion representation must support multiple perception tasks, the model often needs to be redesigned, jointly trained, or repeatedly tuned. This task dependency increases training and deployment costs and limits the cross-task generalisation of the learned fusion representation.

Our work differs from these task-aware fusion methods in that the fusion representation is learned without supervision from a particular downstream task. Instead, we learn a task-agnostic three-channel specialized representation with explicit intensity, gradient, and structural roles. The representation preserves the standard three-channel interface that allows direct compatibility with off-the-shelf downstream networks without modifying their backbones, and can be used to independently train or lightly adapt detection and segmentation networks, requiring only lightweight adaptation of the fusion decoder and retraining. To remain compatible with conventional image-level fusion evaluation, we further introduce a lightweight fusion head after the task-agnostic representation, which is used only to generate a single-channel fused image. In this way, representation learning and task-specific optimisation are separated: the three-channel representation serves downstream perception, while the fusion head supports conventional visualization and image-level evaluation.

3 Method

3.1 Overview

Existing fusion methods typically compress both infrared and visible images into a single fused output using an arbitrary fusion backbone. This single output is expected to satisfy multiple objectives simultaneously: it must preserve thermal saliency and structural details for machine perception (e.g., detection, segmentation), while also appearing visually natural to humans. Optimising these heterogeneous objectives jointly on a single output inevitably leads to trade-offs, where gains in one aspect often degrade another.

To resolve this inherent conflict, as illustrated in Fig. 2, we propose a task-agnostic representation that separates different objectives into dedicated channels. Instead of a single fused image, our framework learns a K-channel representation

$$\mathbf{Z} = F_\phi(I^{\mathrm{ir}}, I^{\mathrm{vi}}) \in [0,1]^{K\times H\times W}, \tag{3}$$

where K is a flexible hyper-parameter: each channel Z_k is supervised by its own dedicated loss $\mathcal{L}_k$, and the number of channels can be increased as new complementary objectives are introduced. In

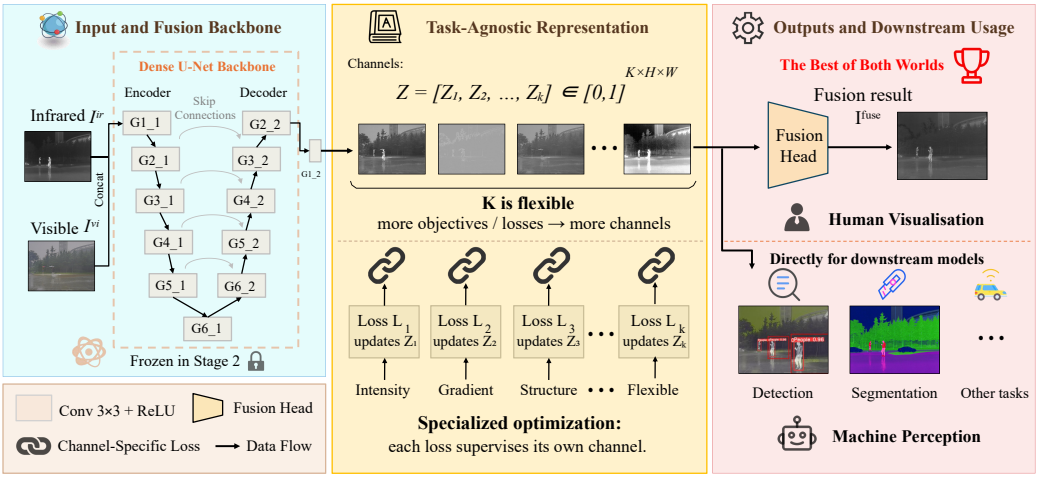

Figure 2: Overview of the proposed task-agnostic representation strategy for infrared–visible image fusion. Our framework adopts a Dense U-Net backbone to learn a K-channel representation, where each channel is supervised by a dedicated loss. Instead of optimising multiple objectives jointly on a single fused output, each loss specializes its corresponding channel separately. The number of channels K is flexible: additional objectives can be introduced by adding more channels. In our instantiation, we set $K = 3$ with explicit intensity, gradient, and structure roles. The resulting representation supports two modes of use: it can be directly fed to standard downstream detection and segmentation networks without architectural changes, and a lightweight fusion head can also be attached to generate a conventional single-channel fused image for visualization and image-level evaluation.

our instantiation, we set $K = 3$, assigning the channels to intensity, gradient, and structural roles. We choose $K = 3$ in this work to match the standard three-channel input format of widely used detection and segmentation networks, allowing our representation to be directly plugged into these off-the-shelf models without any architectural modification.

As illustrated in the center panel of Fig. 2, the learned representation serves two purposes. First, the K-channel representation $\mathbf{Z}$ can be directly fed into off-the-shelf downstream perception networks (e.g., detection, segmentation) without modifying their architectures, since it follows the standard multi-channel input format. Second, a lightweight fusion head H_ψ can project $\mathbf{Z}$ back to a single fused image

$$\hat{I}^{\text{fuse}} = H_\psi(\mathbf{Z}) \in [0,1]^{1\times H\times W}, \tag{4}$$

for conventional image-level evaluation and visualization. These two outputs are optimised in separate stages, described below.

3.2 Stage I: Task-Agnostic Representation Learning

3.2.1 Backbone Architecture

The backbone F_ϕ is implemented as a multiscale encoder–decoder network. Given the concatenated input $\mathbf{X} = [I^{\text{ir}}; I^{\text{vi}}]$, the encoder consists of six convolutional blocks and progressively extracts hierarchical features at multiple scales; a strided convolution at the third block downsamples the feature map to capture a larger receptive field. The decoder then reconstructs the K-channel feature bank $\mathbf{Z} = [Z_1, Z_2, \ldots, Z_K]$ through a symmetric set of convolutional blocks with bicubic upsampling. Following a U-Net-style design, decoder features are concatenated with the corresponding encoder features at each scale, so that fine-grained spatial details are preserved across the downsampling path. All convolutions use 3×3 kernels with reflection padding and ReLU activations, and the channel widths increase from 32 to 64 along the encoder. The final 1×1 projection outputs the K channels followed by a sigmoid activation, yielding $\mathbf{Z} \in [0,1]^{K\times H\times W}$.

3.2.2 CHANNEL-SPECIFIC OBJECTIVES

Each channel of $\mathbf{Z}$ is supervised by a dedicated objective that enforces a distinct aspect of the fused information. Formally, the overall representation loss is the sum of K channel-specific terms:

$$\mathcal{L}_{\text{rep}} = \sum_{k=1}^{K} \mathcal{L}_k(Z_k, I^{\text{ir}}, I^{\text{vi}}), \tag{5}$$

where each $\mathcal{L}_k$ is designed to encourage the k-th channel to focus on one specific aspect of the source images. In our instantiation with $K = 3$, we define three such objectives as follows.

Intensity channel (Z_1). The intensity channel focuses on preserving salient brightness responses from both modalities:

$$\mathcal{L}_1 = \|Z_1 - I^{\text{ir}}\|_2^2 + \|Z_1 - I^{\text{vi}}\|_2^2. \tag{6}$$

Since a thermal target is usually brighter than its surroundings in the infrared modality, this channel retains the cross-modal saliency that object detectors rely on.

Gradient channel (Z_2). The gradient channel focuses on preserving edge structure. We use a fixed high-pass operator $\mathcal{G}(\cdot)$ (a Sobel-style filter) to measure gradient responses and supervise the gradient channel with the element-wise maximum response of the two inputs:

$$\mathcal{L}_2 = \left\|\mathcal{G}(Z_2) - \mathcal{G}\left(\max(I^{\text{ir}}, I^{\text{vi}})\right)\right\|_2^2. \tag{7}$$

Taking the maximum over the two modalities retains the strongest boundaries from either source, which are essential for fine-grained object contours and small-scale structures.

Structure channel (Z_3). The structure channel focuses on preserving the global layout and geometric consistency shared by the two modalities. We adopt the Structural Similarity Index (SSIM) Wang et al. (2004) as the objective for the structure channel. Unlike pixel-wise intensity losses that focus on brightness values, SSIM measures the local luminance, contrast, and structural consistency between two images, which encourages the channel to preserve the overall scene layout and geometric arrangement rather than pixel-level details:

$$\mathcal{L}_3 = \left(1 - \text{SSIM}(Z_3, I^{\text{ir}})\right) + \left(1 - \text{SSIM}(Z_3, I^{\text{vi}})\right). \tag{8}$$

In general, the number of channels K and the choice of objectives are flexible: additional channels can be introduced to capture other complementary aspects of the source images, such as texture or semantic consistency.

3.2.3 OBJECTIVE

Given the frozen representation $\mathbf{Z}$, the head predicts a conventional fused image $\hat{I}^{\text{fuse}}$. It is supervised by a maximum-intensity prior and dual-source structural constraints:

$$\mathcal{L}_{\text{fuse}} = \left\|\hat{I}^{\text{fuse}} - \max(I^{\text{ir}}, I^{\text{vi}})\right\|_1 + \lambda_{\text{s}}\left[\left(1 - \text{SSIM}(\hat{I}^{\text{fuse}}, I^{\text{ir}})\right) + \left(1 - \text{SSIM}(\hat{I}^{\text{fuse}}, I^{\text{vi}})\right)\right], \tag{9}$$

where λ_{s} is a balancing weight (set to 0.5 in our experiments). The ℓ_1 term with the maximum-intensity prior preserves salient thermal targets, while the dual-source SSIM terms retain the structural fidelity of both source images. This stage ensures that the framework remains compatible with conventional image fusion benchmarks while keeping the learned representation unchanged.

3.3 TWO-STAGE TRAINING

Training proceeds in two stages. In the first stage, the backbone F_ϕ is trained alone on paired infrared–visible patches with the representation loss $\mathcal{L}_{\text{rep}}$ in equation 5, using the Adam optimizer with a learning rate of 10^{-4}, a batch size of 5, and 128×128 patches. In the second stage, F_ϕ is frozen and only the fusion head H_ψ is optimised with the fusion loss $\mathcal{L}_{\text{fuse}}$ in equation 9. Freezing the backbone during the second stage is important: it prevents the visualization-oriented objective from perturbing the representation that downstream perception relies on, and keeps the training of the head cheap and stable.

3.4 INFERENCE

At inference time, the task-agnostic representation $\mathbf{Z}$ is directly used as the input to downstream detection or segmentation networks. Since $\mathbf{Z}$ already follows the standard multi-channel format, no change to the downstream architecture is required. If a traditional fused image is needed, the frozen fusion head H_ψ can be applied to $\mathbf{Z}$ to generate $\hat{I}^{\text{fuse}}$ for visualization or image-level evaluation; otherwise, the head can be discarded entirely.

4 EXPERIMENT

4.1 IMPLEMENTATION DETAILS

All experiments are conducted on a Linux server with an NVIDIA GeForce RTX 3090 GPU. Our method takes paired infrared and visible grayscale images as input, with the input channel number set to 2 and the output representation dimension set to 3. We train the model on LLVIP paired patches Cheng et al. (2021) and directly transfer the learned representation to other benchmarks without any fine-tuning. Our model is optimised in two stages. In Stage I, the representation network is trained end-to-end using Adam, outputting a three-channel representation with explicit intensity, gradient, and structure roles. In Stage II, the representation network is frozen and only the lightweight fusion head is trained. Unless otherwise specified, we use a batch size of 5, an initial learning rate of 10^{-4} for both the representation network and the fusion head, weight decay of 10^{-5}, and a random seed of 42. The default number of training epochs is 10 for each stage. The loss weight is set as $\lambda_s = 0.5$.

To evaluate the effectiveness and generalisability of our TAR, we conduct experiments on three public multi-modal fusion benchmarks. All models are trained only on LLVIP and directly tested on other benchmarks, specifically for IVIF we evaluate on LLVIP, MSRS, and TNO Cheng et al. (2026), and for downstream object detection we further test on M³FD Liu et al. (2022), while for semantic segmentation we evaluate on FMB Liu et al. (2023). Our method is compared with seven representative SOTA fusion methods, including LRRNet Li et al. (2023), CrossFuse Li & Wu (2024), GIFNet Cheng et al. (2025), SMLNet Kang et al. (2025), EFRUFN Wang et al. (2026a), ISFuse Wang et al. (2026b), and UniFusion Li et al. (2026). Fusion quality is assessed using four widely adopted metrics, namely EN, MI, VIF and SD Liu et al. (2024). To evaluate downstream perception, we train YOLOv5 Jocher et al. (2022) from scratch on the M³FD detection dataset and SegFormer Xie et al. (2021) on the FMB segmentation dataset.

Table 1: Generalisation study on downstream perception tasks, including M^3FD object detection and FMB semantic segmentation. We apply TAR to two representative fusion backbones: MMDRFuse (ACM MM 2024 Oral, CNN-based) and UPTP (CVPR 2026, Transformer-based), by replacing only the output head. Detection results are reported as $mAP_{50:95}$, and segmentation results as mIoU (%). The improved results after integrating TAR are shown in bold.

Method	M^3FD (Detection)							FMB (Segmentation)						
	People	Car	Bus	Motor	Light	Trunk	mAP	Side	Build	Lamp	Sky	Trunk	Pole	mIoU
MMDRFuse[CNN]	0.460	0.679	0.801	0.441	0.406	0.620	0.568	0.327	0.769	0.045	0.938	0.043	0.386	0.433
+Ours	0.460	**0.680**	**0.822**	**0.457**	0.400	0.614	**0.572**	**0.336**	**0.772**	**0.077**	**0.944**	**0.094**	**0.387**	**0.440**
UPTP[Transformer]	0.459	0.681	0.804	0.434	0.399	0.610	0.564	0.340	0.764	0.080	0.936	0.043	0.381	0.430
+Ours	**0.490**	**0.693**	0.801	**0.459**	**0.429**	**0.640**	**0.585**	**0.412**	**0.780**	**0.110**	**0.942**	**0.106**	**0.402**	**0.490**

4.2 GENERALISATION STUDY

To verify whether the proposed channel-specialization strategy is architecture-agnostic, we apply TAR to two representative fusion networks: MMDRFuse Deng et al. (2024) (CNN-based) and UPTP Deng et al. (2026) (Transformer-based). For each backbone, we replace the conventional single-image output head with our three-channel representation head while keeping the original architecture and training pipeline unchanged. The downstream detection performance on M^3FD ($mAP_{50:95}$) and segmentation performance on FMB (mIoU) are reported in Table 1.

As shown, integrating TAR consistently improves both detection and segmentation performance across both backbones. Notably, every category on FMB is improved after integrating TAR. For

Table 2: Quantitative comparison of the proposed method and existing image fusion methods on the LLVIP, MSRS, and TNO datasets. The best results are in bold.

Method	LLVIP				MSRS				TNO			
	SSIM	VIF	EN	SD	SSIM	VIF	EN	SD	SSIM	VIF	EN	SD
LRRNet	0.6224	0.3421	6.4217	29.8195	0.5871	0.6348	6.1730	31.5221	0.6532	0.6947	6.8734	39.1082
CrossFuse	0.6261	0.3203	6.5229	28.7552	0.6547	0.8133	6.4780	36.0440	0.6523	0.5252	6.7994	37.8925
GIFNet	0.6206	0.7204	7.0012	45.2591	0.6892	0.9833	5.9329	32.3120	0.6599	0.7537	6.7897	39.1731
SMLNet	0.6212	0.3477	6.2289	30.8552	0.6729	0.9283	6.3975	40.1266	0.6381	0.7328	6.9876	42.1286
EFRUFN	0.6650	0.7676	7.2982	47.7576	0.6850	1.1845	6.4887	39.2340	0.6382	0.5588	6.7484	35.4459
ISFuse	0.6237	0.7335	7.1083	43.4709	0.6756	0.9795	6.3832	32.2753	0.6615	0.5038	6.5705	28.7622
UniFusion	0.6303	**0.8912**	7.3951	50.1042	0.6496	1.2350	6.5529	42.3519	**0.6836**	0.6800	6.8404	38.9357
Ours	**0.6720**	0.8652	**7.4303**	**52.2959**	**0.6911**	**1.3166**	**6.6139**	**45.2360**	0.6789	**0.8147**	**7.0341**	**44.9992**

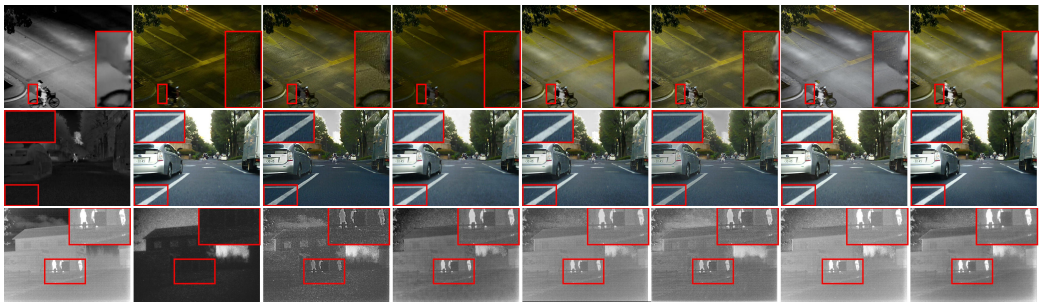

Figure 3: Qualitative results on different datasets (LLVIP, MSRS, and TNO). Our method can maintain an appropriate illumination of the fusion results and preserve abundant texture details.

MMDRFuse (CNN), detection mAP increases from 0.568 to 0.572, and segmentation mIoU from 0.433 to 0.440. The same trend is observed for UPTP (Transformer), where TAR boosts detection mAP from 0.564 to 0.585 and segmentation mIoU from 0.430 to 0.490. These results demonstrate that TAR is not tied to a specific fusion architecture: by simply changing the output from a single fused image to a three-channel representation, existing fusion models can be enhanced without redesigning the backbone, supporting the generality of our approach.

4.3 INFRARED AND VISIBLE IMAGE FUSION

Quantitative Comparisons: Table 2 reports the quantitative comparison. Our method achieves the best overall performance across all three benchmarks, ranking first or second on every evaluation metric. Specifically, our method consistently achieves the strongest VIF, EN, or SD performance while maintaining competitive structural similarity, indicating a favorable balance between structural fidelity and information preservation.

Qualitative Comparisons: We compare visual quality with state-of-the-art methods on LLVIP, MSRS, and TNO datasets (Fig. 3), with red boxes highlighting zoomed-in regions. On LLVIP, GIFNet and SMLNet over-suppress thermal targets, EFRUFN and ISFuse introduce color distortion, while our method keeps the wheel bright with sharp contours. On MSRS, existing methods either produce dark results or over-smooth boundaries, whereas our result preserves clear crossing edges and fine road textures. On TNO, GIFNet and SMLNet fail to balance thermal saliency and details, EFRUFN and ISFuse introduce halo artifacts, while our method maintains strong thermal contrast of the pedestrian while preserving surrounding scene details. These results show that our representation retains complementary information from both modalities and generates informative fused images.

4.4 TASK-DRIVEN DOWNSTREAM IVIF APPLICATIONS

Quantitative Comparisons: To evaluate whether the learned representation is beneficial for machine perception, we conduct downstream object detection and semantic segmentation experiments. Unlike conventional task-aware fusion methods, our fusion representation is learned without detec-

Table 3: Quantitative comparison of our Method and existing image fusion methods for downstream detection and segmentation tasks on the M^{3}FD and FMB datasets.

Method	M^{3}FD (Detection)							FMB (Segmentation)						
	People	Car	Bus	Motor	Light	Trunk	mAP	Side	Build	Lamp	Sky	Trunk	Pole	mIoU
LRRNet	0.450	0.684	0.804	0.462	0.370	0.620	0.565	0.375	0.754	0.045	0.930	0.050	0.379	0.438
CrossFuse	0.438	0.676	**0.813**	0.465	0.408	0.611	0.568	0.348	0.770	0.076	0.931	0.043	0.387	0.418
GIFNet	0.456	0.682	0.803	**0.467**	0.397	0.626	0.572	0.348	0.778	0.065	0.940	0.036	0.397	0.451
SMLNet	0.455	0.673	0.806	0.434	0.413	0.611	0.565	0.335	0.748	0.041	0.925	0.031	0.385	0.418
EFRUFN	0.469	0.684	0.812	0.450	0.401	0.628	0.574	0.357	0.765	0.039	0.940	0.062	0.398	0.435
ISFuse	0.459	0.680	0.800	0.441	0.397	0.619	0.566	0.346	0.779	0.061	0.942	0.050	0.392	0.428
UniFusion	0.464	0.688	0.805	0.450	0.399	**0.635**	0.574	0.341	0.770	0.063	0.937	0.034	0.392	0.427
Ours	**0.481**	**0.691**	0.795	0.463	**0.441**	0.631	**0.584**	**0.385**	**0.782**	**0.158**	**0.947**	**0.112**	**0.418**	**0.452**

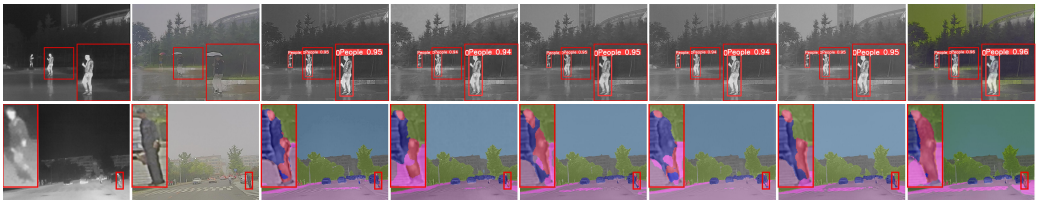

Figure 4: Qualitative results on downstream perception tasks. Our method achieves the highest detection confidence (0.96) on M3FD and produces more complete object masks on FMB compared with state-of-the-art methods.

tion or segmentation supervision, and the resulting three-channel representation is directly used as input to downstream task networks.

Table 3 presents the quantitative results. For object detection, our method achieves the highest mAP$_{50:95}$ of 0.584 on M^{3}FD, outperforming the second-best by 1.0 percentage point. For semantic segmentation, our method obtains the highest mIoU of 0.452 on FMB, ranking first across all six categories. This improvement is consistent with our motivation: while existing methods generate pseudo-RGB inputs without explicitly separating thermal saliency, visible details, and structural information, our three-channel representation provides complementary information through semantically assigned channels.

Qualitative Comparisons: Qualitative downstream results are shown in Fig. 4. For object detection, our method produces more confident detections (confidence score 0.96 vs. 0.94–0.95 for baselines), with bounding boxes more accurately aligned with target regions. For semantic segmentation, our representation produces more complete object masks and sharper boundaries, better separating foreground pedestrians from the background. These results demonstrate that explicitly specialized intensity, gradient, and structural information is more beneficial for downstream perception than a visually fused pseudo-RGB image.

5 Conclusion

This paper presents a task-agnostic three-channel representation for infrared–visible image fusion and downstream perception. By explicitly encoding intensity, gradient, and structural information, the proposed representation reduces the redundancy of pseudo-RGB inputs and provides more informative inputs for standard detection and segmentation networks. A lightweight fusion head is further introduced to generate conventional fused images for image-level evaluation. Experiments on LLVIP, MSRS, TNO, M^{3}FD, and FMB demonstrate that our method achieves strong fusion quality and improves downstream detection and segmentation performance without task-specific fusion optimisation.

6 ACKNOWLEDGMENTS

REFERENCES

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7457–7468. IEEE, 2025.

Chunyang Cheng, Xiao-Jun Wu, Tianyang Xu, and Guoyang Chen. Unifusion: A lightweight unified image fusion network. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for all: Low-level task interaction is a key to task-agnostic image fusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28102–28112. IEEE, 2025.

Chunyang Cheng, Tianyang Xu, Xiao-Jun Wu, Tao Zhou, Hui Li, Zhangyong Tang, and Josef Kittler. Evanet: Towards more efficient and consistent infrared and visible image fusion assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Arthur L Da Cunha, Jianping Zhou, and Minh N Do. The nonsubsampled contourlet transform: theory, design, and applications. *IEEE transactions on image processing*, 15(10):3089–3101, 2006.

Yanglin Deng, Tianyang Xu, Chunyang Cheng, Xiao-Jun Wu, and Josef Kittler. Mmdrfuse: Distilled mini-model with dynamic refresh for multi-modality image fusion. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 7326–7335, 2024.

Yanglin Deng, Tianyang Xu, Chunyang Cheng, Hui Li, Xiaojun Wu, and Josef Kittler. Beyond strict pairing: Arbitrarily paired training for high-performance infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12334–12343, 2026.

Xin Jin, Qian Jiang, Shaowen Yao, Dongming Zhou, Rencan Nie, Jinjin Hai, and Kangjian He. A survey of infrared and visual image fusion methods. *Infrared Physics & Technology*, 85:478–501, 2017.

Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Colin Wong, Zeng Yifu, Diego Montes, et al. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo*, 2022.

Huan Kang, Hui Li, Tianyang Xu, Xiao-Jun Wu, Rui Wang, Chunyang Cheng, and Josef Kittler. Smlnet: A spd manifold learning network for infrared and visible image fusion. *International Journal of Computer Vision*, 133(12):8642–8663, 2025.

Hui Li and Xiao-Jun Wu. Multi-focus image fusion using dictionary learning and low-rank representation. In *International conference on image and graphics*, pp. 675–686. Springer, 2017.

Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE transactions on image processing*, 28(5):2614–2623, 2018.

Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024.

Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):11040–11052, 2023.

Xingyuan Li, Songcheng Du, Yang Zou, HaoYuan Xu, Zhiying Jiang, and Jinyuan Liu. Unifusion: A unified image fusion framework with robust representation and source-aware preservation. *arXiv preprint arXiv:2603.14214*, 2026.

You Li, Julien Moreau, and Javier Ibanez-Guzman. Unconventional visual sensors for autonomous vehicles. *arXiv*, 2022.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5792–5801. IEEE, 2022.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 8081–8090. IEEE, 2023.

Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2226–2235. IEEE, 2025.

Yu Liu, Zhengzheng Qi, Juan Cheng, and Xun Chen. Rethinking the effectiveness of objective evaluation metrics in multi-focus image fusion: A statistic-based approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5806–5819, 2024.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019.

Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.

Oliver Rockinger. Image sequence fusion using a shift-invariant wavelet transform. In *Proceedings of international conference on image processing*, volume 3, pp. 288–291. IEEE, 1997.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 4003–4011, 2022.

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

Alexander Toet. Image fusion by a ratio of low-pass pyramid. *Pattern recognition letters*, 9(4): 245–253, 1989.

Wu Wang, Liang-Jian Deng, Qi Cao, and Gemine Vivone. A general image fusion approach exploiting gradient transfer learning and fusion rule unfolding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a.

Xue Wang, Zheng Guan, Wenhua Qian, Chengchao Wang, and Runzhuo Ma. Multi-modal image fusion via intervention-stable feature learning. *arXiv preprint arXiv:2603.23272*, 2026b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Zengyi Yang, Yu Liu, Juan Cheng, Zhiqin Zhu, Yafei Zhang, and Huafeng Li. Customized fusion: A closed-loop dynamic network for adaptive multi-task-aware infrared-visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 188–198, 2026.

Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 26964–26973. IEEE, 2024.

Qiang Zhang, Yi Liu, Rick S Blum, Jungong Han, and Dacheng Tao. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Information fusion*, 40:57–75, 2018.

Xiaoli Zhang, Liying Wang, Libo Zhao, Xiongfei Li, and Siwei Ma. A semantic-aware and multi-guided network for infrared-visible image fusion. *IEEE Transactions on Multimedia*, 2025.

Xiaoyang Zhang, Jinjiang Li, Guodong Fan, Yakun Ju, Linwei Fan, Jun Liu, and Alex C Kot. Sgdfuse: Sam-guided diffusion model for high-fidelity infrared and visible image fusion. *Information Fusion*, pp. 104290, 2026.

Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020.

Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 13955–13965. IEEE, 2023a.

Wenda Zhao, Wenbo Wang, Haipeng Wang, You He, and Huchuan Lu. Cdtfusion: Crossing domain and task for infrared and visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8048–8059. IEEE, 2023b.

Table 4: Ablation study of channel roles on the M^{3}FD object detection dataset using YOLOv5. All values are reported as fractions of mAP$_{50:95}$. The best results are shown in bold.

Input Representation	M^{3}FD / YOLOv5						Overall mAP$_{50:95}$
	People	Car	Bus	Motor	Light	Truck	
CaseA: Single-channel joint constraint	0.460	0.654	0.793	0.360	0.352	0.613	0.539
CaseB: Grayscale + chrominance	0.473	0.668	0.778	0.423	0.435	0.600	0.563
CaseC01: Intensity-only	0.460	0.682	**0.799**	0.458	0.426	0.621	0.574
CaseC02: Gradient-only	0.465	0.658	0.771	0.360	0.411	0.607	0.545
CaseC03: Structure-only	0.464	0.655	0.781	0.361	0.397	0.590	0.541
Ours	**0.481**	**0.691**	0.795	**0.463**	**0.441**	**0.631**	**0.584**

A Appendix

A.1 Ablation Study

A.1.1 Ablation Study of Channel Roles

To investigate the contribution of different channel roles in the proposed three-channel task-agnostic representation, we conduct an ablation study on the M^{3}FD object detection dataset using YOLOv5. All variants are trained under the same protocol for 200 epochs. Table 4 reports the per-class mAP$_{50:95}$ and the overall mAP$_{50:95}$, where all values are presented as percentages.

We compare several types of input representations. CaseA uses a grayscale fused image learned with a single-channel joint constraint and replicates it three times to construct a three-channel input. CaseB further introduces visible chrominance information into CaseA by combining the grayscale intensity channel with the C_b and C_r chrominance channels. For the role-specific variants, we retain only one of the intensity, gradient, or structure channels and replicate the resulting single-channel map three times to preserve the three-channel input format required by YOLOv5. Finally, Ours denotes the complete task-agnostic representation containing all three channels.

The results are summarized in Table 4. CaseA achieves an overall mAP$_{50:95}$ of only 53.9%, indicating that replicating a single-channel grayscale fused image into three identical channels provides little additional information to the downstream detector. Introducing the C_b and C_r chrominance channels improves the performance of CaseB to 56.3%, suggesting that chrominance information provides useful complementary cues. However, this input still does not explicitly separate the distinct complementary information contained in the infrared and visible modalities.

Among the single-role variants, CaseC01, which retains only the intensity channel, achieves the best performance with an overall mAP$_{50:95}$ of 57.4%, outperforming CaseC02 at 54.5% and CaseC03 at 54.1%. This result highlights the importance of salient intensity information for object detection, particularly for thermally salient targets. Nevertheless, a single channel remains insufficient to fully exploit the complementary information across the two modalities.

The complete Ours representation achieves the best overall mAP$_{50:95}$ of 58.4%. It improves over the single-channel replication baseline by 4.5 percentage points, over the chrominance-augmented CaseB by 2.1 percentage points, and over the strongest single-role variant, CaseC01, by 1.0 percentage point. The complete representation obtains the best results on five of the six object categories, including People, Car, Motorcycle, Lamp, and Truck. It is slightly lower than the intensity-only variant on the Bus category. These results demonstrate that the performance gain does not simply come from increasing the number of input channels. Instead, it results from explicitly assigning complementary functional roles to the intensity, gradient, and structure channels. Together, the three channels provide the downstream detector with complementary thermal saliency, edge details, and structural cues.

A.1.2 Ablation Study on the Number of Channels K

To investigate how the number of specialized channels K affects downstream detection performance, we conduct ablation studies on the M^{3}FD dataset. We consider channel configurations with $K = 1$, $K = 2$, and $K = 3$, where each channel is supervised by one of three dedicated losses: intensity loss

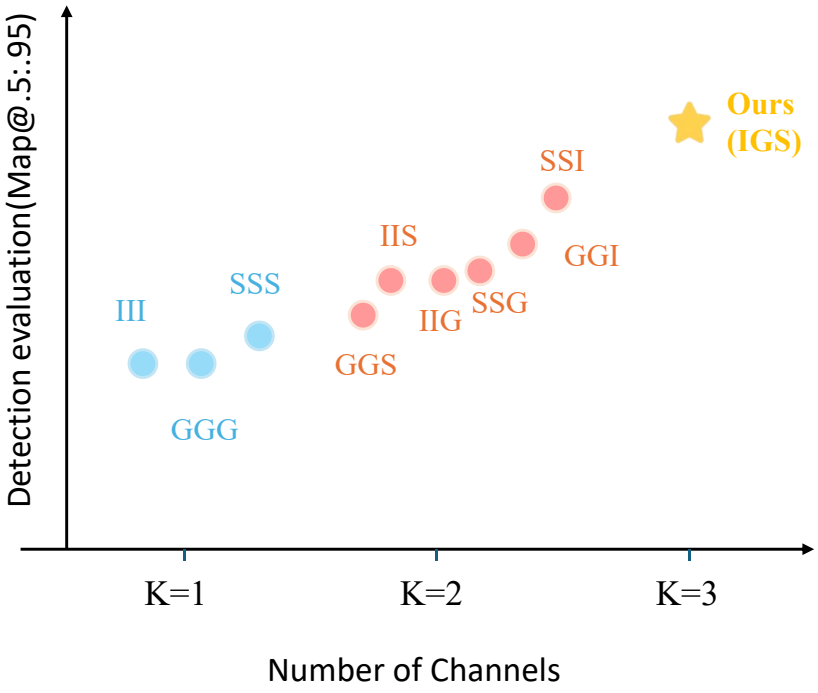

Figure 5: Ablation study on the number of channels K. Each point represents a channel configuration: $K = 1$ shares one objective across all three channels (III, GGG, SSS); $K = 2$ assigns two types of objectives (e.g., IIG, GGS, IIS); $K = 3$ (Ours, IGS) achieves the best detection performance.

(I), gradient loss (G), and structure loss (S). As shown in Fig. 5, when $K = 1$, all three channels share the same objective (III, GGG, or SSS), yielding relatively lower mAP scores. When $K = 2$, the three channels are assigned two types of objectives (e.g., IIG, GGS, IIS), and the detection performance improves to varying degrees depending on the channel combination. When $K = 3$, our full configuration (IGS) assigns intensity, gradient, and structure objectives to separate channels, achieving the best $\text{mAP}_{50:95}$. This trend demonstrates that assigning heterogeneous fusion objectives to dedicated channels consistently improves downstream perception, and our proposed three-channel configuration provides the most informative representation for object detection.When $K > 3$, since standard detection and segmentation networks adopt a fixed three-channel input format, modifying the downstream input architecture would be required. Therefore, we leave the exploration of $K > 3$ configurations as future work.

B LIMITATIONS

Despite the promising results, our method has several limitations that warrant further investigation. First, the three-channel representation is designed for machine perception rather than human visual observation. Although it preserves richer complementary information for downstream networks, its direct RGB visualization may appear less visually natural compared to conventional single-channel fused images, since the three channels encode heterogeneous information rather than a unified grayscale appearance. A lightweight fusion head is required to produce visually plausible outputs for human assessment. Second, the current task-agnostic representation relies on handcrafted loss assignments (intensity, gradient, and structure). Exploring more principled or learned channel decomposition strategies could further improve the representation's expressiveness. Third, our experiments are primarily validated on nighttime driving and surveillance scenarios; extending the representation to broader imaging conditions and downstream tasks remains an interesting direction for future work.