PSDFUSION: PRIOR-GUIDED FEATURE SPECIALIZATION WITH CROSS-PARADIGM DISTILLATION FOR INFRARED-VISIBLE IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared-visible image fusion aims to integrate thermal saliency with visible details while preserving representations useful for downstream perception. Existing prior-guided methods often inject semantic priors into shared multimodal features and transfer teacher knowledge mainly through Euclidean feature matching, which may interfere with low-level visual representation learning and restrict heterogeneous knowledge transfer. We propose PSDFusion, a Prior-guided Specialization and Distillation framework that separates visual specialization from semantic conditioning and enriches student representations across complementary geometries. Specifically, PSDFusion first constructs local-detail and context-preserving representations through asymmetric transformations and cross-gating, followed by text-guided semantic conditioning. It further combines Euclidean and hyperbolic-aware student representations before aligning their channel dependencies with multi-level teacher features in a unified correlation space. Experiments on four infrared-visible fusion benchmarks and two downstream perception tasks against 12 recent baselines demonstrate competitive fusion quality, cross-dataset generalization, and strong task utility. These results support feature specialization and cross-paradigm distillation as complementary mechanisms for balancing visual fidelity and semantic representation learning. The code is publicly available at https://anonymous.4open.science/r/PSDFusion-0CC1.

1 INTRODUCTION

Infrared-visible image fusion (IVIF) aims to integrate thermal saliency from infrared images with fine-grained spatial details from visible images, producing informative representations for robust perception under complex conditions (Gong et al., 2025; Zhou et al., 2025). Beyond improving visual quality, recent studies emphasize preserving task-relevant information for downstream applications such as semantic segmentation and object detection (Tang et al., 2022; Zhao et al., 2025a; Chen et al., 2025; Liu et al., 2025b). Accordingly, high-order priors from vision-language models and pretrained perception networks have been introduced to enhance the semantic utility of fused representations.

For semantic conditioning, infrared and visible features often contain heterogeneous responses associated with fine-grained details, thermal saliency, and broader contextual information. Directly applying high-level semantic priors to such representations may affect visually distinct cues in a similar manner, potentially enhancing task-relevant responses while perturbing local textures, object boundaries, or modality-specific evidence. This indicates that semantic guidance can benefit from a more specialized visual representation before conditioning (Yang et al., 2025; Zhang et al., 2025a; Yi et al., 2024). Accordingly, complementary representation pathways are introduced to emphasize different visual characteristics, providing a more suitable basis for subsequent semantic modulation.

This challenge further extends to knowledge transfer. Teacher features extracted from different network stages exhibit different abstraction characteristics, ranging from local spatial responses to object- and scene-level semantics. Transferring all such knowledge through a single Euclidean representation may restrict the ability of student features to capture these diverse relational patterns. While Euclidean representations naturally preserve local spatial characteristics, alternative geometric views can provide complementary capacity for modeling more complex relationships. We there-

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.

fore construct Euclidean and hyperbolic-aware student representations before distillation, allowing heterogeneous teacher knowledge to be absorbed from complementary representational perspectives.

Taken together, these observations suggest that effective prior-guided fusion depends not only on *what* semantic knowledge is introduced, but also on *how visual representations are prepared for semantic conditioning* and *how heterogeneous teacher knowledge is transferred*. To this end, we propose **PSDFusion**, a **P**rior-guided **S**pecialization and **D**istillation framework for IVIF.

PSDFusion first specializes multimodal features into complementary local-detail and context-preserving representations through asymmetric transformations and reciprocal gating. The former emphasizes fine-grained spatial responses, while the latter retains broader contextual information from the multimodal feature space. Semantic priors are then introduced after feature specialization, allowing high-level guidance to act on a more refined visual basis while reducing unnecessary disturbance to low-level visual cues. Building on the specialized representation, PSDFusion further constructs complementary Euclidean and hyperbolic-aware student views for knowledge distillation. The Euclidean pathway preserves the native spatial characteristics of student features, whereas the hyperbolic-aware pathway provides an additional nonlinear representation for relational modeling. The two views are subsequently integrated and aligned with multi-level teacher features in a unified correlation space. This progressive design enables semantic guidance and teacher supervision to operate on representations tailored to their respective roles, improving semantic representation learning while preserving visual fidelity. Our main contributions are summarized as follows:

- We propose **PSDFusion**, a prior-guided specialization and distillation framework that progressively adapts multimodal representations for semantic conditioning and heterogeneous knowledge transfer across diverse visual representation spaces.
- We develop a prior-guided feature specialization mechanism that constructs complementary local-detail and context-preserving representations through asymmetric transformations and reciprocal gating, providing a refined visual basis for subsequent semantic conditioning.
- We introduce cross-paradigm distillation that combines complementary Euclidean and hyperbolic-aware student representations before multi-level teacher supervision, enabling more flexible knowledge transfer while preserving visual fidelity.

2 RELATED WORK

2.1 MULTIMODAL ALIGNMENT AND FEATURE SPECIALIZATION

Modern IVIF methods have gradually shifted from direct pixel-level aggregation toward representation-level alignment and complementary feature modeling. TDFusion (Bai et al., 2025) introduces a learnable task-driven objective to preserve perception-relevant information. DDBFusion (Zhang et al., 2025b) and FreeFusion (Zhao et al., 2025b) exploit dual decomposition and cross-reconstruction, respectively, to retain modality-specific characteristics during feature interaction. AWFusion (Li et al., 2026) adopts adaptive weighting to improve robustness under adverse conditions, while CrossFuse (Shi et al., 2025b) promotes cross-modal interaction through multi-view alignment. MulFS-CAP (Li et al., 2025) further incorporates spatial-aware interaction to reduce geometric discrepancies across modalities.

Despite these advances, existing methods mainly focus on how infrared and visible information is aligned, reconstructed, or adaptively combined. When high-level semantic guidance is subsequently introduced, shared multimodal features may still contain heterogeneous responses related to local details, thermal saliency, and broader contextual information. Applying semantic modulation directly to such representations may affect visually distinct cues simultaneously, potentially weakening fine-grained details or modality-specific evidence.

2.2 HIGH-ORDER SEMANTIC PRIORS AND DISTILLATION

Recent IVIF methods increasingly exploit high-order knowledge from generative models (Ho et al., 2020), Vision-Language Models (VLMs) (Li et al., 2023), and pretrained perception networks (Xie et al., 2021). VDMUFusion (Shi et al., 2025a) and Mask-DiFuser (Tang et al., 2026) employ diffusion-based modeling to capture contextual dependencies, while PromptFusion (Liu et al., 2025a) and

OmniFuse (Zhang et al., 2025a) introduce natural-language guidance to enhance semantic awareness. ZVIR (Zhai et al., 2026) further incorporates deep image priors under structural constraints. These studies demonstrate the value of external knowledge beyond conventional low-level fusion objectives.

However, existing studies focus primarily on how external priors are introduced, while the representation used to absorb and transfer such knowledge receives less attention. Features extracted from different stages of pretrained perception networks exhibit heterogeneous abstraction characteristics, ranging from local spatial responses to high-level semantics. Conventional distillation typically transfers these features within a Euclidean representation space, which preserves native spatial correspondence but may provide limited flexibility for modeling diverse relational patterns across levels. Hyperbolic-aware representations offer an additional nonlinear view, complementing the spatial characteristics retained by Euclidean features.

3 METHOD

Figure 1 illustrates the PSDFusion architecture. PSDFusion addresses prior-guided IVIF from two perspectives. First, *prior-guided feature specialization* transforms multimodal features into complementary local-detail and context-preserving pathways before semantic conditioning, enabling high-level guidance to act on a refined visual basis without perturbing entangled features. Second, *cross-paradigm distillation* constructs Euclidean and hyperbolic-aware student representations before transferring multi-level teacher knowledge. Together, these components improve semantic representation learning while preserving complementary multimodal source information.

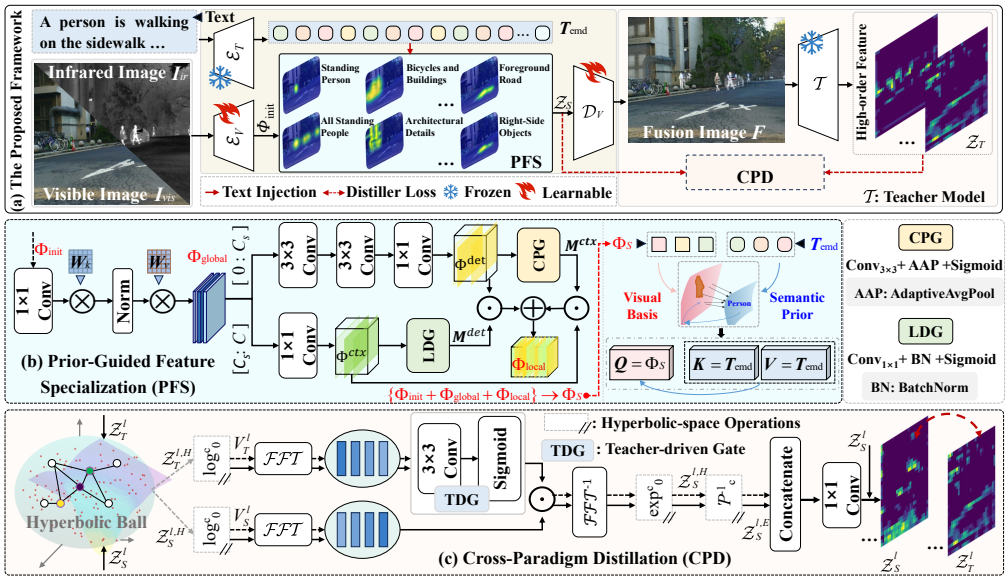

Figure 1: Overview of PSDFusion. (a) Overall framework. (b) Feature specialization constructs local-detail and context-preserving representations before semantic conditioning. (c) Cross-paradigm distillation integrates Euclidean and hyperbolic-aware representations for teacher supervision.

3.1 FORMULATION

Given an infrared image I_{ir}, a visible image I_{vis}, and a textual condition Text, PSDFusion learns a fusion network $\mathcal{F}_\Theta$ that produces the fused image F and multi-level student representations:

$$\left\{F, \{Z_{\mathrm{S}}^l\}_{l\in\mathcal{L}}\right\} = \mathcal{F}_\Theta\left(I_{\mathrm{ir}}, I_{\mathrm{vis}}, \mathrm{Text}\right), \tag{1}$$

where Θ denotes the learnable parameters and $l \in \mathcal{L}$ indexes the representation levels.

To provide perception-oriented supervision, a pretrained SegFormer is adopted as a frozen teacher $\mathcal{T}$. Given F, it extracts multi-level teacher representations:

$$\{Z_{\mathrm{T}}^l\}_{l\in\mathcal{L}} = \mathcal{T}(F), \tag{2}$$

which capture complementary information across different abstraction levels.

The fusion network is optimized by

$$\Theta^* = \arg\min_{\Theta} \left[\mathcal{L}_{\mathrm{f}} + \lambda \mathcal{L}_{\mathrm{d}} \left(\{\widetilde{Z}_{\mathrm{S}}^l\}_{l\in\mathcal{L}}, \{Z_{\mathrm{T}}^l\}_{l\in\mathcal{L}} \right) \right], \tag{3}$$

where $\mathcal{L}_{\mathrm{f}}$ preserves source visual information, $\mathcal{L}_{\mathrm{d}}$ transfers multi-level teacher knowledge, and $\widetilde{Z}_{\mathrm{S}}^l$ denotes the student representation refined by cross-paradigm distillation.

3.2 Prior-Guided Feature Specialization

Directly applying high-level semantic priors to shared multimodal features may disturb fine-grained details and modality-specific cues. We therefore introduce **Prior-Guided Feature Specialization (PFS)**, which constructs complementary visual representations with different inductive biases before semantic conditioning. This specialization-first strategy provides a refined visual basis for high-level guidance, allowing semantic priors to interact with multimodal features more selectively while preserving complementary visual information.

Initial Multimodal Representation. Given the infrared and visible images, $\mathcal{E}_{\mathrm{V}}(\cdot)$ combines 3×3 convolutions with Sobel and Laplacian operators to capture appearance and gradient information:

$$\Phi_{\mathrm{init}} = \mathcal{E}_{\mathrm{V}} \left(I_{\mathrm{ir}} \oplus I_{\mathrm{vis}} \right), \tag{4}$$

where $\oplus$ denotes channel-wise concatenation. To complement local visual responses with broader contextual information, Φ_{init} is projected into a globally aggregated representation:

$$\Phi_{\mathrm{global}} = \mathrm{Norm} \left(\Phi_{\mathrm{init}} W_k^{\mathrm{T}} \right) W_v, \tag{5}$$

where W_k and W_v are learnable projection matrices. The resulting representation captures global dependencies while retaining the multimodal information encoded in Φ_{init}.

Complementary Feature Specialization. Although aggregation facilitates contextual modeling, it may smooth fine-grained variations. We therefore partition Φ_{global} along the channel dimension at C_s, where $0 < C_s < C$, and process both components using asymmetric transformations:

$$\begin{aligned} \Phi_{\mathrm{det}} &= \mathrm{Conv}_{3\times3} \left(\mathrm{Conv}_{3\times3} \left(\mathrm{Conv}_{1\times1} \left(\Phi_{\mathrm{global}}[0 : C_s] \right) \right) \right), \\ \Phi_{\mathrm{ctx}} &= \mathrm{Conv}_{1\times1} \left(\Phi_{\mathrm{global}}[C_s : C] \right), \end{aligned} \tag{6}$$

where C denotes the total number of channels. The deeper transformations on Φ_{det} emphasize spatially varying and fine-grained responses, whereas the lightweight transformation on Φ_{ctx} preserves globally aggregated contextual information with less spatial perturbation. Accordingly, we regard them as *local-detail* and *context-preserving* representations, respectively. Rather than assigning explicit semantic roles, the asymmetric transformations introduce complementary inductive biases that facilitate cross-branch interaction and semantic conditioning.

To promote information exchange between the specialized pathways, we further employ reciprocal cross-branch gating:

$$M_{\mathrm{det}} = \mathrm{LDG} \left(\Phi_{\mathrm{ctx}} \right), \qquad M_{\mathrm{ctx}} = \mathrm{CPG} \left(\Phi_{\mathrm{det}} \right), \tag{7}$$

where LDG and CPG denote Local-Detail Gating and Context-Preserving Gating, respectively. The gates adaptively reweight each pathway using information from its counterpart, enabling contextual information to refine local responses and cues to modulate contextual features.

The gated responses are aggregated as

$$\Phi_{\mathrm{local}} = M_{\mathrm{det}} \odot \Phi_{\mathrm{det}} + M_{\mathrm{ctx}} \odot \Phi_{\mathrm{ctx}}, \tag{8}$$

where $\odot$ denotes element-wise multiplication. The initial, global, and specialized representations are then integrated through residual aggregation:

$$\Phi_{\mathrm{S}} = \Phi_{\mathrm{init}} + \Phi_{\mathrm{global}} + \Phi_{\mathrm{local}}, \tag{9}$$

yielding a refined multimodal representation for subsequent semantic conditioning.

Semantic Conditioning after Specialization. After visual specialization, the textual condition is encoded as T_{emd} by a pretrained BLIP-2(Li et al., 2023). Semantic guidance is then introduced through cross-modal interaction with Φ_{S}:

$$Z_{\mathrm{S}} = \Phi_{\mathrm{S}} + \mathrm{Softmax}\left(Q_{\Phi_{\mathrm{S}}} K_{T_{\mathrm{emd}}}^{\mathrm{T}}\right) V_{T_{\mathrm{emd}}}, \tag{10}$$

where $Q_{\Phi_{\mathrm{S}}}$ is projected from visual representation, and $K_{T_{\mathrm{emd}}}$ and $V_{T_{\mathrm{emd}}}$ from textual representation.

The resulting Z_{S} preserves the visual complementarity established during specialization while incorporating high-level semantic guidance. Finally, the decoder reconstructs the fused image:

$$F = \mathcal{D}_{\mathrm{V}}\left(Z_{\mathrm{S}}\right), \tag{11}$$

where $\mathcal{D}_{\mathrm{V}}(\cdot)$ contains three 3×3 Conv–LReLU blocks followed by a 3×3 Conv–Tanh layer.

3.3 Cross-Paradigm Distillation

The frozen teacher provides multi-level representations with varying abstraction characteristics, which may be difficult to transfer through a single representation geometry. We therefore introduce **Cross-Paradigm Distillation (CPD)**, which constructs complementary Euclidean and hyperbolic-aware student representations: the former preserves native spatial characteristics, while the latter provides a nonlinear view for relational modeling. These representations are integrated with the original student feature and distilled in a unified correlation space.

Hyperbolic-Aware Representation Refinement. For the l-th level, the student feature Z_{S}^{l} and dimension-aligned teacher feature $\mathrm{Align}(Z_{\mathrm{T}}^{l})$ are mapped to the Poincaré ball (Nickel & Kiela, 2017):

$$Z_{\mathrm{S}}^{l,\mathrm{H}} = \mathcal{P}_c\left(Z_{\mathrm{S}}^{l}\right), \qquad Z_{\mathrm{T}}^{l,\mathrm{H}} = \mathcal{P}_c\left(\mathrm{Align}\left(Z_{\mathrm{T}}^{l}\right)\right), \tag{12}$$

where $\mathcal{P}_c(\cdot)$ denotes the mapping from Euclidean space to the Poincaré ball with curvature magnitude c, and $\mathrm{Align}(\cdot)$ matches the spatial and channel dimensions of teacher and student features.

Since Fourier operations are performed in Euclidean space, both representations are projected to the tangent space at the origin:

$$V_k^{l} = \log_0^c\left(Z_k^{l,\mathrm{H}}\right), \qquad k \in \{\mathrm{S}, \mathrm{T}\}, \tag{13}$$

where $\log_0^c(\cdot)$ denotes the logarithmic map. Teacher-guided spectral interaction is then performed in the tangent space:

$$\begin{aligned} \Gamma_{\mathrm{T}}^{l,\mathrm{H}} &= \mathrm{TDG}\left(\mathrm{FFT}\left(V_{\mathrm{T}}^{l}\right)\right), \\ \widehat{Z}_{\mathrm{S}}^{l,\mathrm{H}} &= \exp_0^c\left[\mathrm{FFT}^{-1}\left(\mathrm{FFT}\left(V_{\mathrm{S}}^{l}\right) \odot \Gamma_{\mathrm{T}}^{l,\mathrm{H}}\right)\right], \end{aligned} \tag{14}$$

where $\mathrm{TDG}(\cdot)$ denotes teacher-driven gating and $\exp_0^c(\cdot)$ maps the tangent representation to the Poincaré ball. $\widehat{Z}_{\mathrm{S}}^{l,\mathrm{H}}$ provides a teacher-conditioned hyperbolic-aware view of student features.

Euclidean Representation Refinement. In parallel, teacher-guided spectral interaction is performed in the original Euclidean space:

$$\begin{aligned} \Gamma_{\mathrm{T}}^{l,\mathrm{E}} &= \mathrm{TDG}\left(\mathrm{FFT}\left(\mathrm{Align}\left(Z_{\mathrm{T}}^{l}\right)\right)\right), \\ \widehat{Z}_{\mathrm{S}}^{l,\mathrm{E}} &= \mathrm{FFT}^{-1}\left(\mathrm{FFT}\left(Z_{\mathrm{S}}^{l}\right) \odot \Gamma_{\mathrm{T}}^{l,\mathrm{E}}\right). \end{aligned} \tag{15}$$

This pathway retains the native Euclidean representation of the student while incorporating teacher-conditioned spectral responses. Together, the two pathways provide complementary representation views for subsequent knowledge transfer.

Cross-Paradigm Representation Integration. The hyperbolic-aware representation is first mapped back to Euclidean space and concatenated with its Euclidean counterpart:

$$U_{\mathrm{S}}^{l} = \mathrm{Concat}\left[\widehat{Z}_{\mathrm{S}}^{l,\mathrm{E}}, \mathcal{P}_c^{-1}\left(\widehat{Z}_{\mathrm{S}}^{l,\mathrm{H}}\right)\right]. \tag{16}$$

The combined representation is integrated with the student feature through a residual connection:

$$\widetilde{Z}_{\mathrm{S}}^{l} = Z_{\mathrm{S}}^{l} + \mathrm{Conv}_{1\times1}\left(U_{\mathrm{S}}^{l}\right). \tag{17}$$

This residual formulation preserves the original response while incorporating complementary information from both geometric views.

Correlation-Based Knowledge Distillation. After cross-paradigm refinement, teacher supervision is imposed in a common Euclidean correlation space:

$$\mathcal{L}_{\mathrm{d}}^{l} = \left\| \mathcal{G}\left(\widetilde{Z}_{\mathrm{S}}^{l}\right) - \mathcal{G}\left(\mathrm{Align}\left(Z_{\mathrm{T}}^{l}\right)\right) \right\|_2^2, \quad \text{s.t.} \quad \mathcal{G}(Z) = \mathrm{Norm}_2\left(\bar{Z}\bar{Z}^{\mathrm{T}}\right). \tag{18}$$

Here, $\bar{Z}$ denotes the flattened representation and $\mathrm{Norm}_2(\cdot)$ denotes row-wise ℓ_2 normalization.

3.4 Training Objective

To preserve complementary source information, we adopt intensity and gradient constraints:

$$\mathcal{L}_{\mathrm{int}} = \left\| F - \max(I_{\mathrm{ir}}, I_{\mathrm{vis}}) \right\|_1, \mathcal{L}_{\mathrm{grad}} \quad = \left\| \nabla F - \max(\nabla I_{\mathrm{ir}}, \nabla I_{\mathrm{vis}}) \right\|_1. \tag{19}$$

The overall objective combines fusion fidelity with multi-level teacher distillation:

$$\mathcal{L} = \mathcal{L}_{\mathrm{int}} + \mathcal{L}_{\mathrm{grad}} + \lambda \sum_{l\in\mathcal{L}} \mathcal{L}_{\mathrm{d}}^{l}, \tag{20}$$

where λ balances fusion fidelity and teacher supervision.

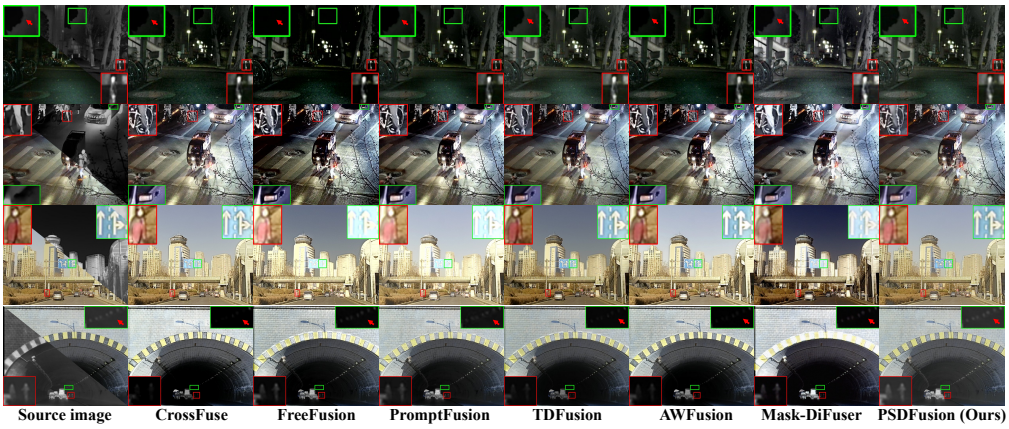

Figure 2: Qualitative comparisons between Ours and state-of-the-art methods. From top to bottom, the representative examples are selected from the MSRS, LLVIP, FMB, and M3FD datasets, respectively.

4 Experiments

4.1 Experimental Setup

Datasets and Benchmarks PSDFusion is trained on the MSRS (Ha et al., 2017) training split (1,083 pairs) and evaluated on its test split (361 pairs), with textual descriptions from VLFDataset (Zhao et al., 2024). For cross-dataset evaluation, the trained model is directly tested on LLVIP (Jia et al., 2021) (300 pairs), FMB (Liu et al., 2023) (280 pairs), and M3FD (Liu et al., 2022) (300 pairs) without further training. As scene-level annotations are unavailable, fixed prompts "This is an infrared image." and "This is a visible image." are used. We compare with 12 representative methods: CrossFuse (Shi et al., 2025b), DDBFusion (Zhang et al., 2025b), FreeFusion (Zhao et al., 2025b), PromptFusion (Liu et al., 2025a), MulFS-CAP (Li et al., 2025), OmniFuse (Zhang et al., 2025a), TDFusion (Bai et al., 2025), TextFusion (Cheng et al., 2025), VDMUFusion (Shi et al., 2025a), AWFusion (Li et al., 2026), Mask-DiFuser (Tang et al., 2026), and ZVIR (Zhai et al., 2026). We evaluate fusion quality using AG, MI (Qu et al., 2002), $Q_{ab/f}$ (Xydeas et al., 2000), IFC (Sheikh et al., 2005), ED (Abdou & Pratt, 1979), and Q_{cb} (Chen & Blum, 2009), where higher values indicate better performance.

Table 1: Quantitative comparison with 12 representative methods on the MSRS and LLVIP datasets. The best and second-best results are highlighted in **bold** and **bold underline**, respectively.

Pub.	Methods	MSRS						LLVIP					
		AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
TCSVT'25	CrossFuse	5.134	2.490	0.493	1.912	46.66	0.449	**8.021**	2.443	0.548	1.981	**62.12**	0.504
INF'25	DDBFusion	3.538	2.191	0.376	1.948	29.24	0.495	4.858	2.458	0.437	2.020	37.52	0.490
TPAMI'25	FreeFusion	4.415	1.984	0.421	1.575	39.90	0.400	7.897	2.361	0.472	1.771	60.98	0.430
JAS'25	PromptFusion	4.446	3.619	0.553	2.560	41.11	0.464	5.134	2.109	0.372	1.466	40.07	0.485
TPAMI'25	MulFS-CAP	3.037	1.396	0.251	0.546	31.38	0.404	3.526	1.225	0.124	0.283	35.74	0.389
TPAMI'25	OmniFuse	3.210	2.844	0.401	1.491	33.06	0.440	4.792	2.221	0.284	1.375	46.47	0.422
CVPR'25	TDFusion	4.335	**3.741**	**0.667**	**3.932**	39.64	**0.548**	6.961	**3.322**	**0.659**	**3.127**	56.31	**0.531**
INF'25	TextFusion	3.459	3.347	0.472	2.031	32.60	0.484	3.640	2.238	0.276	1.436	27.93	0.443
TIP'25	VDMUFusion	2.905	2.268	0.333	2.110	22.82	0.499	4.438	2.705	0.382	2.211	32.77	0.479
INF'26	AWFusion	4.268	2.801	0.601	3.088	39.20	0.525	6.297	2.809	0.594	2.542	52.09	0.500
TPAMI'26	Mask-DiFuser	**5.664**	2.382	0.511	2.324	**48.11**	0.478	**8.280**	2.852	0.573	2.308	**65.08**	0.498
PR'26	ZVIR	4.347	2.497	0.515	2.106	40.85	0.510	5.441	2.138	0.360	1.507	45.21	0.473
–	**Ours**	**5.883**	**4.671**	**0.711**	**4.838**	**52.00**	**0.590**	6.550	**3.526**	**0.651**	**3.233**	55.98	**0.529**

Table 2: Quantitative comparison with 12 representative methods on the FMB and M3FD datasets. The best and second-best results are highlighted in **bold** and **bold underline**, respectively.

Pub.	Methods	FMB						M3FD					
		AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
TCSVT'25	CrossFuse	6.737	3.832	0.675	3.585	54.59	**0.470**	9.337	3.303	0.506	2.113	72.84	0.463
INF'25	DDBFusion	3.784	2.964	0.431	2.175	31.08	0.416	5.529	2.675	0.430	2.325	41.52	0.423
TPAMI'25	FreeFusion	**7.819**	3.056	0.588	2.838	**64.79**	0.469	**10.12**	2.815	0.539	2.800	**76.84**	**0.474**
JAS'25	PromptFusion	4.878	**4.476**	**0.681**	**7.681**	42.26	0.460	6.922	**3.898**	0.576	3.269	52.53	0.446
TPAMI'25	MulFS-CAP	4.140	2.582	0.416	0.974	35.70	0.396	5.746	2.308	0.395	1.051	46.43	0.411
TPAMI'25	OmniFuse	2.979	2.659	0.310	1.460	29.13	0.451	3.078	2.846	0.271	1.432	31.09	0.453
CVPR'25	TDFusion	5.708	3.365	0.676	3.654	47.63	0.469	7.171	3.170	**0.629**	**3.469**	55.31	**0.486**
INF'25	TextFusion	5.907	3.278	0.565	2.233	48.74	0.449	8.198	3.027	0.472	2.041	63.06	0.436
TIP'25	VDMUFusion	3.314	3.029	0.390	2.427	26.32	0.454	4.360	2.817	0.368	2.416	32.11	0.440
INF'26	AWFusion	6.199	3.155	0.657	3.473	51.98	0.448	7.425	3.049	0.608	3.354	57.86	0.462
TPAMI'26	Mask-DiFuser	**8.364**	3.120	0.480	2.494	**68.23**	0.398	**10.09**	2.828	0.444	2.485	**77.12**	0.439
PR'26	ZVIR	5.137	2.749	0.539	2.189	42.32	0.427	7.489	2.430	0.480	2.116	58.96	0.444
–	**Ours**	5.876	**4.475**	**0.689**	**4.397**	49.98	**0.476**	7.428	**4.334**	**0.621**	**3.919**	56.30	0.460

Implementation Details PSDFusion is implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU using Adam for 100 epochs. The batch size and learning rate are 6 and 5×10^{-4}, respectively. The curvature, distillation weight, and number of representation levels are set to $c = 1$, $\lambda = 0.5$, and $L = 3$. Additional sensitivity and ablation analyses are provided in the **Appendix**.

4.2 Comparison with State-of-the-Art Methods

Qualitative Comparison. Fig. 2 presents qualitative comparisons on the four benchmarks. In low-light MSRS and LLVIP scenes, several competing methods, such as CrossFuse and AWFusion, produce blurred boundaries or weakened local textures. In comparison, PSDFusion preserves clearer target contours and finer structural details, as highlighted by the red boxes. For the more complex FMB and M3FD scenes, FreeFusion and PromptFusion occasionally introduce local overexposure or color distortion, whereas PSDFusion maintains salient infrared targets while retaining visible-background structures. These results show that PSDFusion provides a favorable balance between thermal saliency preservation and fine-grained visual detail.

Quantitative Comparison. Tables 1 and 2 report quantitative results on the four benchmarks. On MSRS, PSDFusion achieves the best performance across all six metrics, demonstrating strong information and structural preservation on the training-domain benchmark. For cross-dataset evaluation, PSDFusion remains competitive despite using generic modality prompts. Specifically, it achieves the best MI and IFC on LLVIP and M3FD, the best $Q_{ab/f}$ and Q_{cb} on FMB, while obtaining second-best results on several complementary metrics. Although detail-sensitive measures such as AG and ED

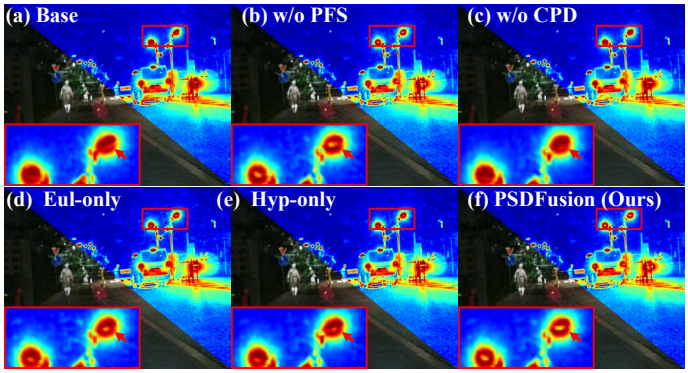

Figure 3: Qualitative analysis under different architectures.

Table 3: Quantitative ablation results of the proposed method. The best and second-best results are highlighted in **bold** and **bold underline**, respectively.

Setting	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
Base	4.221	4.653	0.697	4.780	39.51	0.567
w/o PFS	5.604	**4.834**	0.710	**6.005**	49.89	**0.590**
w/o CPD	5.649	4.703	0.708	**5.510**	50.10	0.584
Eul-only	**5.859**	**4.700**	**0.712**	4.948	**51.92**	**0.588**
Hyp-only	5.768	4.697	0.707	5.149	51.23	0.585
Ours	**5.883**	4.671	**0.711**	4.838	**52.00**	**0.590**

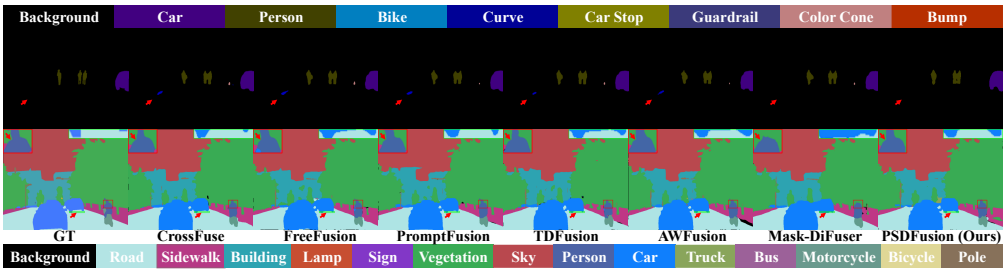

Figure 4: Qualitative semantic segmentation comparison on MSRS and FMB datasets.



Figure 5: Qualitative object detection comparison on LLVIP and M3FD datasets.

are not consistently dominant across all datasets, the overall results indicate stable generalization and a balanced preservation of source information, structural fidelity, and perceptual quality.

4.3 Ablation Study

We evaluate five variants: a baseline without feature specialization and cross-paradigm distillation (**Base**), two component ablations (**w/o PFS** and **w/o CPD**), and two single-geometry variants (**Eul-**

Table 4: Quantitative comparison of downstream tasks across four benchmark datasets. The best and second-best results are highlighted in **bold** and **bold underline**, respectively.

Methods	Semantic Segmentation (mIoU / %)						Object Detection (mAP / %)							
	MSRS			FMB			LLVIP				M3FD			
	Crv.	Guar.	mIoU	Road	Buil.	mIoU	P	R	@.5	@.5:.95	P	R	@.5	@.5:.95
CrossFuse	43.88	49.12	65.20	88.73	**82.24**	53.05	**94.98**	**86.56**	93.39	57.13	85.82	70.79	78.64	53.82
DDBFusion	44.90	45.66	66.74	89.11	81.65	52.70	92.82	85.22	91.65	57.45	84.58	76.01	81.23	54.97
FreeFusion	29.78	55.72	62.18	88.52	80.99	53.16	94.08	83.35	91.74	57.29	84.54	**81.42**	**85.13**	58.43
PromptFusion	46.60	60.74	68.82	88.97	81.05	52.52	90.27	85.46	92.56	58.14	**87.60**	74.94	82.21	56.59
MulFS-CAP	39.91	59.12	59.93	87.42	78.77	50.06	86.71	75.49	83.06	41.68	84.07	74.99	79.97	52.61
OmniFuse	44.04	58.95	67.18	88.40	79.66	52.39	89.57	71.87	82.82	48.60	85.51	72.40	80.16	54.75
TDFusion	**48.83**	63.46	**69.81**	88.51	80.39	52.92	92.67	84.12	91.47	58.22	87.12	81.05	84.95	**59.39**
TextFusion	41.74	**71.12**	68.00	86.63	75.29	49.72	92.36	86.27	**93.55**	**59.61**	83.76	77.35	82.03	55.44
VDMUFusion	45.11	63.80	66.47	89.00	81.90	53.39	92.15	85.37	91.26	56.59	87.42	74.99	80.94	56.08
AWFusion	39.33	64.48	66.60	88.31	81.53	**53.65**	93.64	83.98	91.33	56.37	82.77	80.65	84.16	59.23
Mask-DiFuser	47.70	46.98	67.67	89.16	81.75	53.46	94.39	83.81	91.48	57.38	87.18	78.41	83.80	57.87
ZVIR	47.25	65.07	69.70	**89.20**	80.80	52.50	93.15	79.95	89.66	56.91	**88.46**	75.51	82.96	57.24
Ours	**54.17**	**74.30**	**71.37**	**89.36**	**82.28**	**54.18**	**94.42**	**86.66**	**93.80**	**61.63**	86.56	**81.39**	**85.14**	**59.44**

only and **Hyp-only**). As shown in Fig. 3, Base and w/o PFS lose fine-grained details, while removing CPD leads to weaker structural preservation. The single-geometry variants show less complete structures than the full model. Table 3 reports the quantitative results. Base achieves the weakest performance, demonstrating the benefit of feature specialization and cross-paradigm refinement. Removing PFS improves MI and IFC but decreases AG and ED, indicating that feature specialization benefits detail and edge preservation rather than uniformly improving all metrics. Eul-only yields a slightly higher $Q_{ab/f}$, whereas combining Euclidean and hyperbolic-aware representations achieves a more balanced trade-off across structural, information, and perceptual measures.

4.4 Downstream Perception Tasks

For downstream evaluation, we adopt SegFormer (Xie et al., 2021) for semantic segmentation and YOLOv8 (Varghese & M., 2024) for object detection, using identical settings for all methods.

For semantic segmentation, Fig. 4 shows that competing methods produce incomplete boundaries in challenging regions, whereas PSDFusion retains clearer contours on MSRS. On FMB, it better preserves structures around objects such as wheels and persons, as highlighted by the marked regions. For object detection, Fig. 5 shows that PSDFusion maintains recognizable target structures under low illumination and partial occlusion, reducing localization errors in LLVIP and M3FD.

Quantitative results are reported in Table 4. PSDFusion achieves mIoU scores of 71.37% on MSRS and 54.18% on FMB, together with strong performance on challenging categories such as Crv., Guar., Road, and Buil. For object detection, it obtains the best mAP@0.5:0.95 on LLVIP and M3FD, reaching 61.63% and 59.44%, respectively, while remaining competitive in precision, recall, and mAP@0.5. These results indicate that the fused representations preserve task-relevant information beyond conventional image-level fusion quality.

5 Conclusion

We present PSDFusion, a prior-guided feature specialization and cross-paradigm distillation framework for infrared-visible image fusion. PSDFusion constructs complementary local-detail and context-preserving representations before semantic conditioning, and further enriches student features through Euclidean and hyperbolic-aware transformations for multi-level teacher supervision in a unified correlation space. Experiments on four fusion benchmarks and two downstream perception tasks demonstrate competitive fusion quality, cross-dataset generalization, and strong task utility. Ablation studies further verify the complementary contributions of PFS and CPD components. Despite these advantages, cross-paradigm transformations introduce additional training overhead, while static or generic prompts may limit scene-adaptive semantic guidance. Future work will explore lightweight geometric representations and adaptive vision-language prompting to improve efficiency and semantic generalization.

AI USE STATEMENT

Generative AI tools were used only for language polishing to improve grammar, clarity, and conciseness. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Ikram E Abdou and William K Pratt. Quantitative design and evaluation of enhancement/thresholding edge detectors. *Proceedings of the IEEE*, 67(5):753–763, 1979. doi: 10.1109/PROC.1979.11325.

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 International Conference on Computer Vision (CVPR)*, pp. 7457–7468, Nashville, TN, USA, 2025. IEEE. doi: 10.1109/CVPR52734.2025.00699.

Jun Chen, Liling Yang, Wei Yu, Wenping Gong, Zhanchuan Cai, and Jiayi Ma. Sdsfusion: A semantic-aware infrared and visible image fusion network for degraded scenes. *IEEE Transactions on Image Processing*, 34:3139–3153, 2025. doi: 10.1109/TIP.2025.3571339.

Yin Chen and Rick S. Blum. A new automated quality assessment algorithm for image fusion. *Image and Vision Computing*, 27(10):1421–1432, 2009. ISSN 0262-8856. doi: 10.1016/j.imavis.2007.12.002.

Chunyang Cheng, Tianyang Xu, Xiao-Jun Wu, Hui Li, Xi Li, Zhangyong Tang, and Josef Kittler. Textfusion: Unveiling the power of textual semantics for controllable image fusion. *Information Fusion*, 117:102790, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102790.

Meiqi Gong, Hao Zhang, Xunpeng Yi, Linfeng Tang, and Jiayi Ma. Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In *2025 IEEE/CVF International Conference on Computer Vision (CVPR)*, pp. 14326–14335. IEEE, 2025.

Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115, Vancouver, BC, Canada, 2017. IEEE. doi: 10.1109/IROS.2017.8206396.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *2020 International Conference on Neural Information Processing Systems (NIPS)*, pp. 6840–6851, Red Hook, NY, USA, 2020. Curran Associates Inc. doi: 10.5555/3495724.3496298.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3489–3497, Montreal, BC, Canada, 2021. IEEE. doi: https://doi.org/10.1109/ICCVW54120.2021.00389.

Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3673–3690, 2025. doi: 10.1109/TPAMI.2025.3535617.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *2023 International Conference on Machine Learning (ICML)*, pp. 19730–19742, Honolulu, HI, USA, 2023. PMLR. doi: https://doi.org/10.5555/3618408.3619222.

Xilai Li, Wuyang Liu, Xiaosong Li, Fuqiang Zhou, Huafeng Li, and Feiping Nie. All-weather multi-modality image fusion: Unified framework and 100k benchmark. *Information Fusion*, 131:104130, 2026. doi: https://doi.org/10.1016/j.inffus.2026.104130.

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, New Orleans, LA, USA, 2022. IEEE. doi: 10.1109/CVPR52688.2022.00571.

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, Paris, France, 2023. IEEE. doi: 10.1109/ICCV51070.2023.00745.

Jinyuan Liu, Xingyuan Li, Zirui Wang, Zhiying Jiang, Wei Zhong, Wei Fan, and Bin Xu. PromptFusion: Harmonized Semantic Prompt Learning for Infrared and Visible Image Fusion. *IEEE/CAA Journal of Automatica Sinica*, 12(3):502–515, 2025a. doi: https://doi.org/10.1109/JAS.2024.124878.

Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025b. doi: 10.1109/TPAMI.2024.3521416.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *2017 International Conference on Neural Information Processing Systems (NIPS)*, pp. 6341–6350, Long Beach, CA, USA, 2017. Curran Associates, Inc. doi: 10.5555/3295222.3295381.

Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):1, 2002. doi: 10.1049/EL:20020212.

Hamid Rahim Sheikh, Alan Conrad Bovik, and Gustavo de Veciana de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005. doi: 10.1109/TIP.2005.859389.

Yu Shi, Yu Liu, Juan Cheng, Z. Jane Wang, and Xun Chen. Vdmufusion: A versatile diffusion model-based unsupervised framework for image fusion. *IEEE Transactions on Image Processing*, 34:441–454, 2025a. doi: 10.1109/TIP.2024.3512365.

Yukai Shi, Cidan Shi, Zhipeng Weng, Yin Tian, Xiaoyu Xian, and Liang Lin. Crossfuse: Learning infrared and visible image fusion by cross-sensor top-k vision alignment and beyond. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(8):7579–7591, 2025b. doi: 10.1109/TCSVT.2025.3544746.

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. doi: https://doi.org/10.1016/j.inffus.2021.12.004.

Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion network with language-vision degradation prompts. In *2025 International Conference on Neural Information Processing Systems (NIPS)*. Curran Associates Inc., 2025.

Linfeng Tang, Chunyu Li, and Jiayi Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):591–608, 2026. doi: 10.1109/TPAMI.2025.3609323.

Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6, Chennai, India, 2024. doi: https://doi.org/10.1109/ADICS58448.2024.10533619.

Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17882–17891, Nashville, TN, USA, 2025. IEEE. doi: 10.1109/CVPR52734.2025.01666.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: simple and efficient design for semantic segmentation with transformers. In *2021 International Conference on Neural Information Processing Systems (NIPS)*, NIPS '21, pp. 12077–12090, Red Hook, NY, USA, 2021. Curran Associates Inc. doi: https://doi.org/10.5555/3540261.3541185.

Costas S Xydeas, Vladimir Petrovic, and others. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000. doi: 10.1049/el:20000267.

Chenjia Yang, Xiaoqing Luo, Zhancheng Zhang, Zhiguo Chen, and Xiao jun Wu. Kdfuse: A high-level vision task-driven infrared and visible image fusion method based on cross-domain knowledge distillation. *Information Fusion*, 118:102944, 2025. doi: https://doi.org/10.1016/j.inffus.2025.102944.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27016–27025, Seattle, WA, USA, 2024. IEEE. doi: 10.1109/CVPR52733.2024.02552.

Hao Zhai, Minjie Deng, Yan Wei, Long Wang, and Anyu Li. Zvir: Zero-shot implicit deep image prior with prior activation for infrared and visible image fusion. *Pattern Recognition*, 172:112666, 2026. doi: https://doi.org/10.1016/j.patcog.2025.112666. URL https://www.sciencedirect.com/science/article/pii/S0031320325013299.

Hao Zhang, Lei Cao, Xuhui Zuo, Zhenfeng Shao, and Jiayi Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025a. doi: 10.1109/TPAMI.2025.3568433.

Zeyang Zhang, Hui Li, Tianyang Xu, Xiao-Jun Wu, and Josef Kittler. DDBFusion: An unified image decomposition and fusion framework based on dual decomposition and Bézier curves. *Information Fusion*, 114:102655, 2025b. doi: https://doi.org/10.1016/j.inffus.2024.102655.

Jiayi Zhao, Fei Teng, Kai Luo, Guoqiang Zhao, Zhiyong Li, Xu Zheng, and Kailun Yang. Unveiling the potential of segment anything model 2 for rgb-thermal semantic segmentation with language guidance. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2193–2200, Hangzhou, China, 2025a. IEEE. doi: https://doi.org/10.1109/IROS60139.2025.11247337.

Wenda Zhao, Hengshuai Cui, Haipeng Wang, You He, and Huchuan Lu. Freefusion: Infrared and visible image fusion via cross reconstruction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8040–8056, 2025b. doi: 10.1109/TPAMI.2025.3572599.

Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, and Luc Van Gool. Image fusion via vision-language model. In *2024 International Conference on Machine Learning (ICML)*, pp. 60749 – 60765, Vienna, Austria, 21–27 Jul 2024. PMLR. doi: doi/10.5555/3692070.3694583.

Man Zhou, Jie Huang, Keyu Yan, Danfeng Hong, Xiuping Jia, Jocelyn Chanussot, and Chongyi Li. A general spatial-frequency learning framework for multimodal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5281–5298, 2025. doi: 10.1109/TPAMI.2024.3368112.

A Appendix

A.1 Extended Ablation Studies

Effect of the distillation weight λ. The weight λ balances multi-level teacher supervision and image-level fidelity. As shown in Table 5, a small value such as $\lambda = 0.1$ provides limited distillation guidance, whereas increasing λ generally improves structural and perceptual quality. Although $\lambda = 0.3$ achieves the highest MI, its performance on the remaining metrics is less competitive. Further increasing λ to 0.7 or 1.0 places greater emphasis on distillation and weakens the overall balance across fusion metrics. Therefore, $\lambda = 0.5$ is adopted as the default setting, providing the most balanced performance across the evaluated measures.

Table 5: Ablation study on the balance hyperparameter λ. The best and second-best results are highlighted in **bold** and underline, respectively.

λ	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
0.1	5.779	4.708	0.697	4.749	50.75	0.566
0.3	5.831	**4.732**	0.696	4.698	51.07	0.570
0.5	**5.883**	4.671	**0.711**	**4.838**	**52.00**	**0.590**
0.7	5.826	4.575	0.686	4.615	50.78	0.570
1.0	5.849	4.389	0.677	4.337	50.81	0.562

Table 6: Ablation study on the curvature hyperparameter c.

c	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
0.1	5.871	4.587	0.690	4.440	51.32	0.572
0.5	5.828	4.744	0.693	**4.846**	50.93	0.570
1.0	5.883	4.671	**0.711**	4.838	**52.00**	**0.590**
1.5	**5.884**	4.738	0.692	4.766	51.43	0.573
2.0	5.858	**4.777**	0.693	4.836	51.18	0.571

Effect of the curvature magnitude c. The curvature magnitude c controls the geometric transformation strength of the hyperbolic-aware pathway in CPD. As shown in Table 6, $c = 0.1$ yields weaker overall performance, while larger values such as 1.5 and 2.0 lead to less balanced results across the evaluation metrics. Although $c = 0.5$ achieves a relatively high IFC score, its performance on structural measures remains less competitive. Therefore, $c = 1.0$ is adopted as the default setting, providing the best overall balance across the evaluated metrics.

Table 7: Ablation study on the number of distillation levels L.

L	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
1	5.306	4.420	0.643	3.875	47.05	0.554
3	**5.883**	4.671	**0.711**	**4.838**	52.00	**0.590**
5	5.826	4.330	0.667	3.861	51.73	0.536
7	5.832	**4.786**	0.692	4.833	51.05	0.572
9	5.843	3.792	0.609	3.424	**52.06**	0.529

Effect of the number of distillation levels L. The number of distillation levels L controls how many teacher–student feature pairs participate in the multi-level correlation supervision of CPD. As shown in Table 7, $L = 1$ provides limited multi-level supervision and yields weaker overall performance. Increasing L incorporates teacher information from different abstraction levels, but further increasing the number of levels does not consistently improve all metrics. Although $L = 7$ achieves the highest MI, it is less competitive on structural and perceptual measures. Therefore, $L = 3$ is adopted as the default setting, providing the most balanced overall performance.

Effect of the channel split point C_s. The split point C_s controls the channel allocation between the local-detail and context-preserving pathways in PFS, where $C_s = C/\alpha$ and $\alpha \in \{2, 4, 6, 8, 10\}$. As shown in Table 8, $\alpha = 2$ allocates more channels to the local-detail pathway and achieves the highest MI and IFC, but performs less consistently on the remaining metrics. As α increases, fewer channels are assigned to the local-detail pathway and more to the context-preserving pathway. The settings $\alpha = 6$ and $\alpha = 8$ yield weaker overall results, although $\alpha = 8$ remains competitive in AG and ED. With $\alpha = 10$, detail-related performance remains relatively strong, but the overall results are still inferior to those of $\alpha = 4$. Therefore, $\alpha = 4$, corresponding to $C_s = C/4$, is adopted as the default setting for its balanced performance across AG, $Q_{ab/f}$, ED, and Q_{cb}.

Table 8: Ablation study on the channel split point C_s, where $C_s = C/\alpha$.

α	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
2	5.724	**4.783**	0.694	**4.909**	50.31	0.574
4	**5.883**	4.671	**0.711**	4.838	**52.00**	**0.590**
6	5.830	4.591	0.690	4.476	51.12	0.568
8	5.872	4.581	0.689	4.543	51.27	0.573
10	5.874	4.580	0.692	4.487	51.31	0.573

Table 9: Textual conditions used in the ablation study.

Condition	**Textual Input**
Empty	No textual input.
Task-oriented prompt	"A high-quality fused image combining infrared and visible light information. The goal is to merge thermal radiation targets with visible spatial details for enhanced scene representation."
Modality-level	"This is an infrared image." "This is a visible image."
Scene-level	Paired infrared and visible captions describing the specific scene content.

Table 10: Ablation results under different textual conditions.

Condition	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
Empty	5.355	4.813	0.699	4.921	50.39	0.578
Task-oriented	5.362	5.648	0.696	4.547	50.32	0.575
Modality-level	5.085	**5.026**	0.696	**5.778**	48.33	0.569
Scene-level	**5.883**	4.671	**0.711**	4.838	**52.00**	**0.590**

Effect of different textual conditions. Table 9 summarizes four textual conditions with different semantic granularity and annotation requirements: empty input, a fixed task-oriented instruction, modality-level prompts, and paired scene-level captions. This experiment examines how PSDFusion responds to textual guidance ranging from generic task information to sample-specific scene descriptions. Scene-level captions provide instance-specific semantic context, whereas modality-level prompts require no scene annotations and only identify the input modalities.

The quantitative results are reported in Table 10. The fixed task-oriented instruction provides general fusion guidance but limited sample-specific information. Modality-level prompts achieve the highest MI and IFC, indicating that simple modality identity already provides useful guidance for preserving source information without additional annotations. Paired scene-level captions achieve the best AG, $Q_{ab/f}$, ED, and Q_{cb}, suggesting that finer-grained semantic descriptions can further benefit structural and perceptual preservation. Accordingly, scene-level captions are used on MSRS when such annotations are available, while fixed modality prompts are adopted for annotation-free cross-dataset evaluation on FMB, LLVIP, and M3FD. Overall, PSDFusion benefits from richer textual context while remaining effective under coarse, annotation-free modality guidance.

B Dense Prediction Teacher

We further evaluate the choice of frozen teacher in CPD by replacing SegFormer with DeepLabV3+, UPerNet, and BiSeNetV2 under identical settings. As shown in Table 11, SegFormer achieves the best overall performance across the six fusion metrics, indicating that its multi-level representations provide effective supervision for correlation-based distillation. Despite architectural differences

Table 11: Ablation of the dense prediction teacher under identical training settings. Only the teacher architecture is changed, while all other configurations remain fixed.

Teacher	AG	MI	$Q_{ab/f}$	IFC	ED	Q_{cb}
DeepLabV3+	5.332	<u>4.595</u>	<u>0.690</u>	4.596	<u>50.02</u>	<u>0.570</u>
UPerNet	<u>5.392</u>	4.551	0.689	<u>4.648</u>	49.81	0.568
BiSeNetV2	5.352	4.494	0.686	4.470	49.93	<u>0.570</u>
SegFormer	**5.883**	**4.671**	**0.711**	**4.838**	**52.00**	**0.590**

among the teachers, PSDFusion maintains competitive performance across all settings, showing that CPD is not tied to a particular dense prediction backbone. SegFormer is therefore adopted as the default teacher due to its more balanced overall performance, while the results also demonstrate the robustness of PSDFusion to different teacher architectures.

C Complexity Analysis

Table 12: Comparison of computational complexity in terms of parameter count, FLOPs, and inference time. The best and second-best results are highlighted in **bold** and <u>underline</u>, respectively.

Method	**Params. (M)**	**FLOPs (G)**	**Time (s)**
CrossFuse	1.1900	547.74	**0.0226**
DDBFusion	10.920	369.90	0.7249
FreeFusion	5.6653	451.65	0.0942
PromptFusion	0.3600	111.40	0.3759
MulFS-CAP	0.3750	**7.7500**	6.2884
OmniFuse	2.0100	56.910	6.9030
TDFusion	**0.0593**	36.422	<u>0.0386</u>
TextFusion	<u>0.0688</u>	<u>20.794</u>	0.0974
VDMUFusion	31.751	63.253	46.060
AWFusion	19.360	4167.2	0.9240
Mask-DiFuser	171.26	4667.4	3.0314
ZVIR	7.9200	62.070	8.2200
Ours	3.1210	627.85	0.4067

Table 12 compares model size, FLOPs, and inference time across different fusion methods. PSDFusion contains 3.121M parameters and requires 0.4067 s per image, placing its model size and runtime in the middle range of the compared methods. Although slower than several lightweight baselines, it remains faster than seven competing methods, including diffusion- and iterative-based approaches. Its computational cost reaches 627.85 GFLOPs, the third highest among the compared methods, mainly due to dense feature processing involving global aggregation, complementary local-detail and context-preserving pathways, cross-branch interaction, semantic conditioning, and convolutional reconstruction. Despite the relatively high FLOPs, PSDFusion maintains moderate inference latency under the reported setting. Overall, PSDFusion is not designed as a lightweight model, but provides a practical balance among parameter size, inference efficiency, fusion quality, and downstream perception performance.

D Additional Comparison Results

Quantitative Comparison Figs. 6–9 compare six fusion metrics across four datasets. PSDFusion achieves the best results on all metrics for MSRS and remains competitive on the three cross-dataset benchmarks. Specifically, it obtains the best MI and IFC on LLVIP and M3FD, and the best $Q_{ab/f}$ and Q_{cb} on FMB, while remaining competitive on several complementary metrics. Although AG and ED are not consistently dominant on unseen datasets, the overall results indicate a balanced preservation

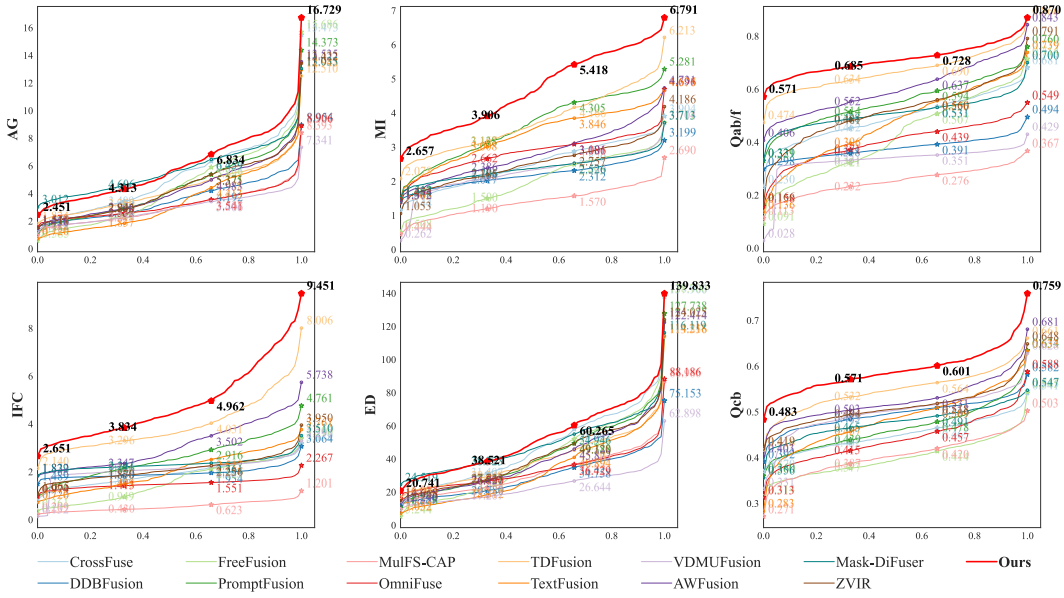

Figure 6: Quantitative comparison of PSDFusion with 12 methods on MSRS.

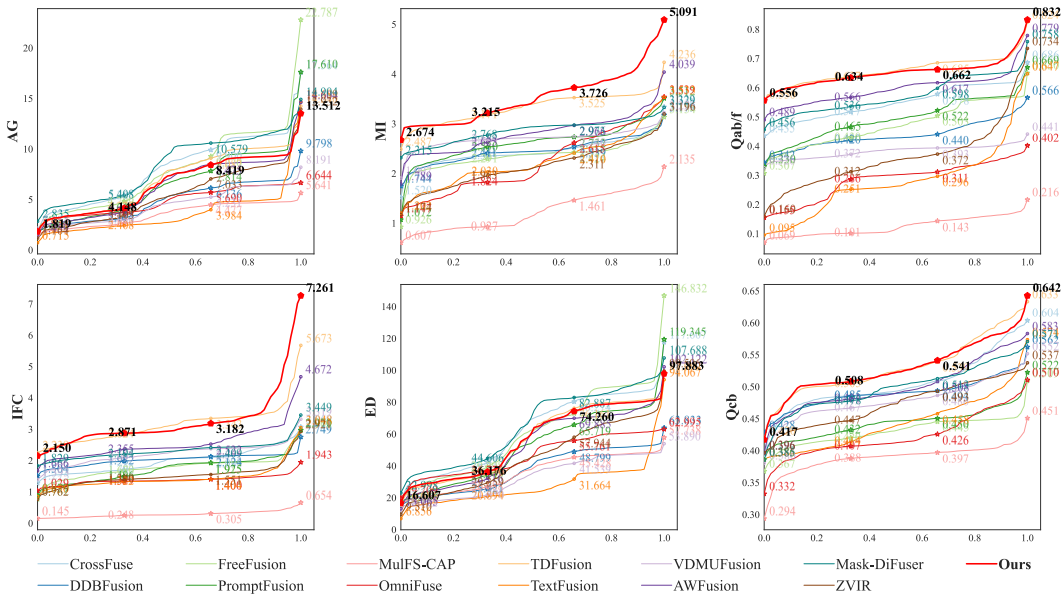

Figure 7: Quantitative comparison of PSDFusion with 12 methods on LLVIP.

of source information, structural fidelity, and perceptual quality. These results further support the cross-dataset generalization of PSDFusion, particularly on information- and structure-related metrics.

Qualitative Comparison As shown in Fig. 10, MSRS contains scenes with substantial illumination variations, posing challenges to simultaneous preservation of target boundaries and background details. CrossFuse and ZVIR exhibit blurred contours in bright regions, while DDBFusion loses local boundary information under low illumination. In comparison, PSDFusion preserves clearer target contours and more consistent structural details across different lighting conditions.

For nighttime scenes in LLVIP, Fig. 11 shows that FreeFusion and AWFusion weaken small structures such as headlights and local object boundaries. PSDFusion retains salient infrared responses while

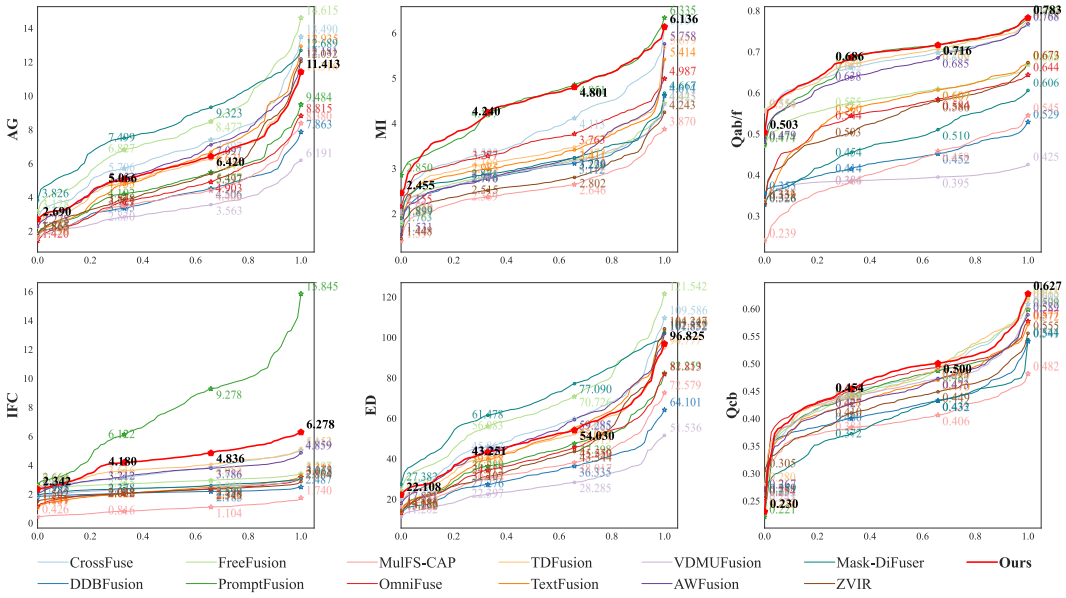

Figure 8: Quantitative comparison of PSDFusion with 12 methods on FMB.

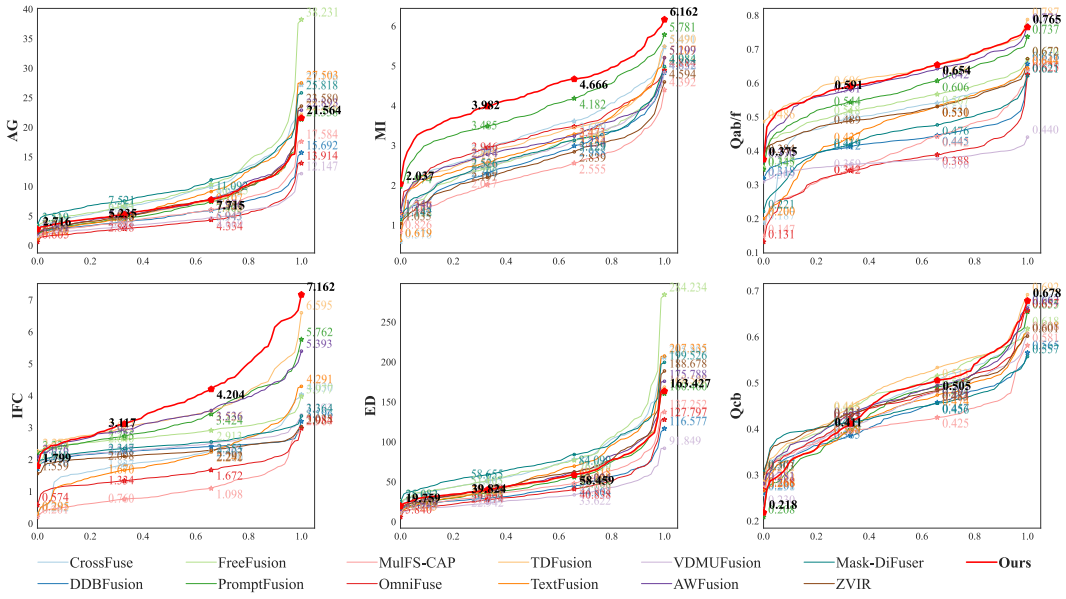

Figure 9: Quantitative comparison of PSDFusion with 12 methods on M3FD.

preserving visible textures in surrounding regions, yielding a more balanced representation of foreground targets and background information.

As illustrated in Fig. 12, FMB contains scenes with uneven illumination and complex color distributions. DDBFusion produces relatively blurred structures, while FreeFusion and Mask-DiFuser introduce local overexposure or color shifts in several regions. PSDFusion maintains clearer structures and more stable contrast while reducing severe local artifacts, resulting in a more natural visual appearance.

Fig. 13 further compares the methods in complex urban and tunnel scenes. PromptFusion, TextFusion, and VDMUFusion exhibit weakened object boundaries or unbalanced preservation between salient

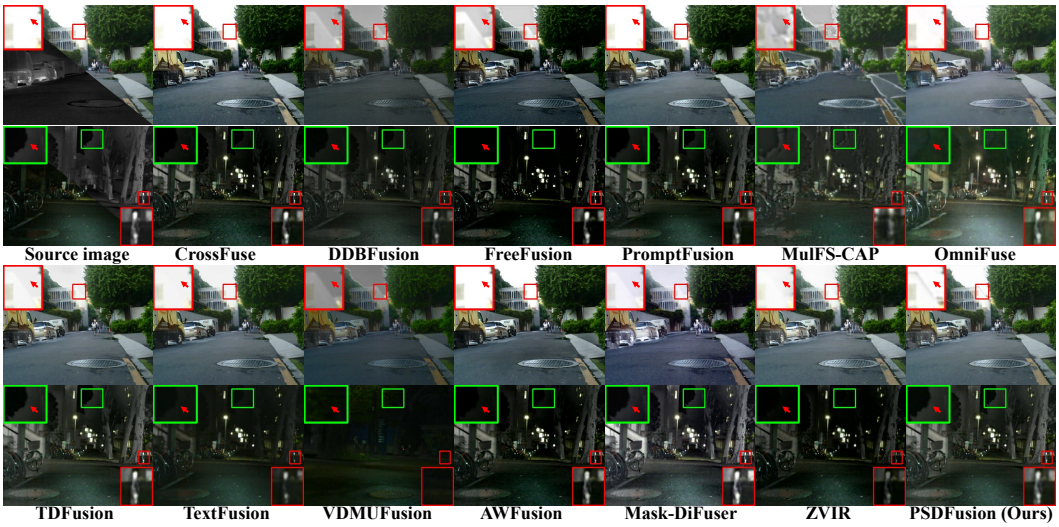

Figure 10: Qualitative comparison of PSDFusion with 12 methods on MSRS.



Figure 11: Qualitative comparison of PSDFusion with 12 methods on LLVIP.

targets and strong light sources. By contrast, PSDFusion retains more complete target contours and surrounding scene details while preserving prominent infrared responses.

E Extended Validation on High-level Vision Tasks

Semantic segmentation. Fig. 14 presents representative semantic segmentation results on MSRS. In complex regions, CrossFuse and AWFusion produce fragmented predictions or inaccurate category boundaries, as indicated by the red arrows. Predictions based on PSDFusion images retain more continuous regions and clearer boundaries. In the magnified areas, MulFS-CAP and AWFusion exhibit noticeable boundary deviations around objects such as *Car*, whereas PSDFusion better preserves the visual cues required for accurate region delineation.

Fig. 15 further compares segmentation results on FMB. For foreground instances, PSDFusion yields more continuous person contours in the red boxes and preserves finer local structures, such as wheels, in the green boxes. For background categories such as *Building*, the predicted regions are also more spatially consistent with fewer isolated errors. Several distant or small targets that are missed or

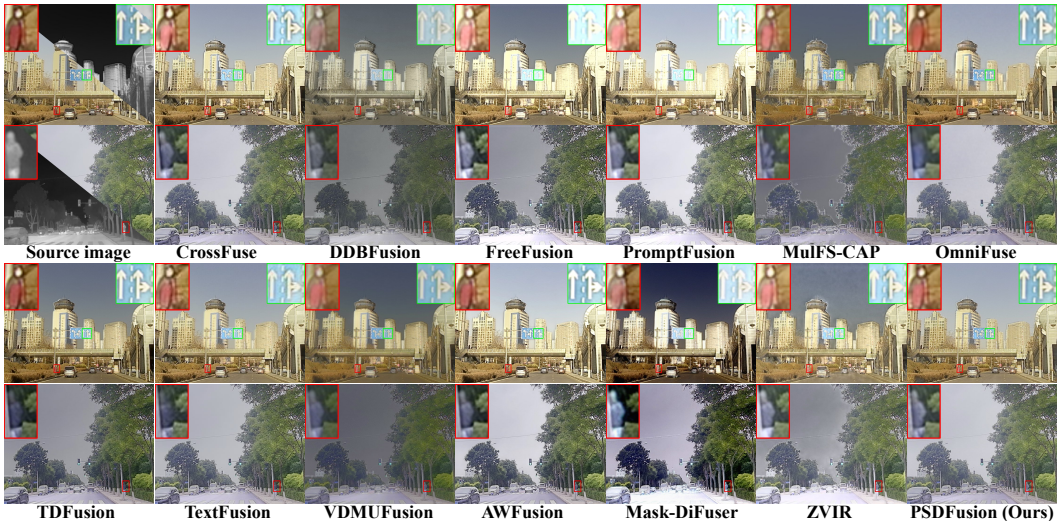

Figure 12: Qualitative comparison of PSDFusion with 12 methods on FMB.

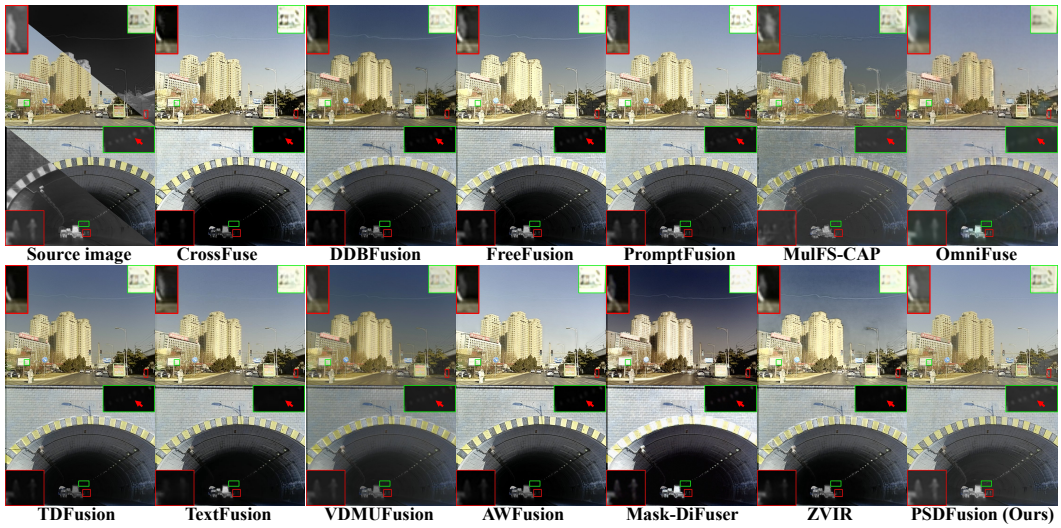

Figure 13: Qualitative comparison of PSDFusion with 12 methods on M3FD.

incompletely segmented by competing methods are also better identified from PSDFusion images. These observations indicate that PSDFusion preserves visual information beneficial to downstream semantic segmentation.

Object detection. Fig. 16 shows representative object detection results on LLVIP, where heavy occlusion, small targets, and low illumination make pedestrian detection challenging. AWFusion and Mask-DiFuser miss several partially occluded persons or produce less accurate bounding boxes. In comparison, PSDFusion enables the detector to localize more small or partially visible instances. In dim and cluttered street regions, DDBFusion and ZVIR also exhibit missed detections or larger localization deviations, whereas PSDFusion produces more accurate detections in the displayed examples.

The M3FD examples in Fig. 17 contain overlapping objects and cluttered urban backgrounds, which further challenge target localization. FreeFusion misses several *Car* instances, while OmniFuse produces inaccurate predictions for some foreground objects. The detector operating on PSDFusion images retains more complete detections and better-localized bounding boxes for the illustrated

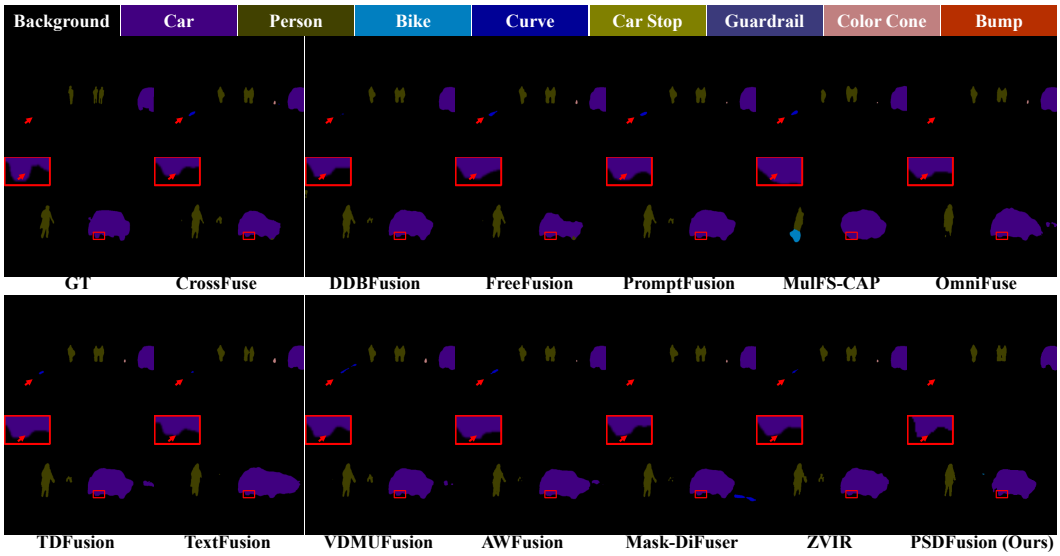

Figure 14: Qualitative comparison of semantic segmentation performance between the proposed PSDFusion and 12 state-of-the-art methods on MSRS.

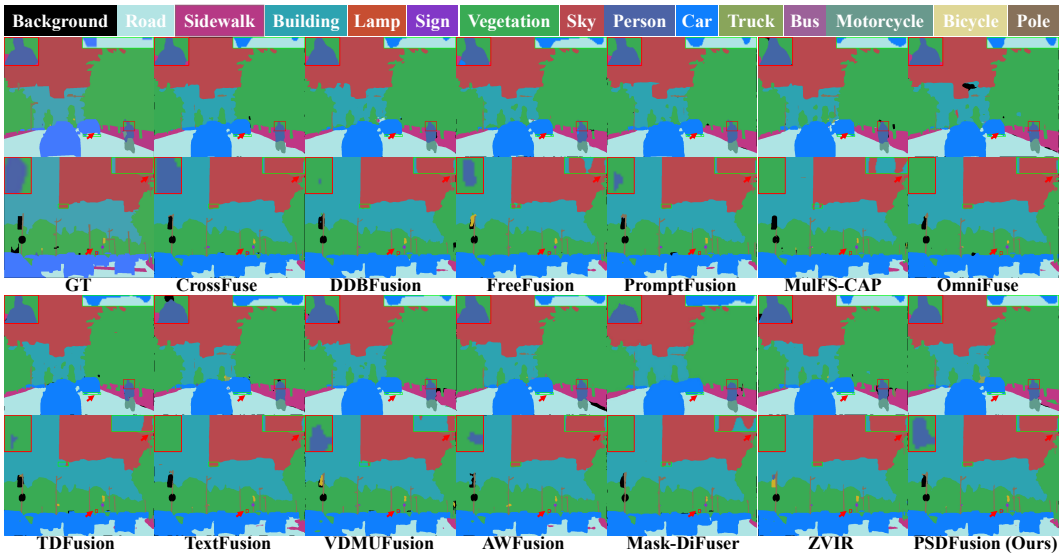

Figure 15: Qualitative comparison of semantic segmentation performance between the proposed PSDFusion and 12 state-of-the-art methods on FMB.

Person and *Car* instances. Overall, these qualitative results indicate that PSDFusion preserves target and structural cues useful for downstream detection under occlusion, low illumination, and complex backgrounds.

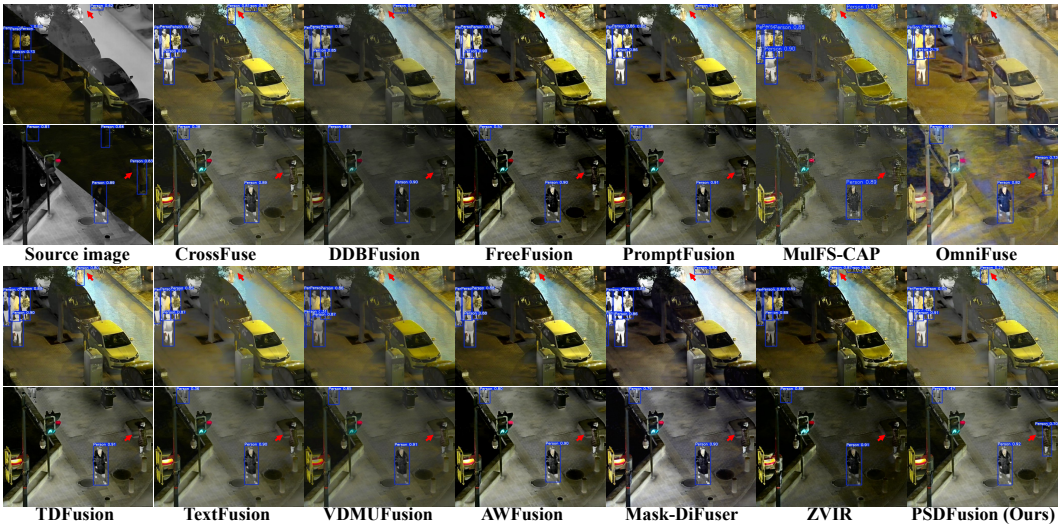

Figure 16: Qualitative comparison of object detection performance between the proposed PSDFusion and 12 state-of-the-art methods on LLVIP.

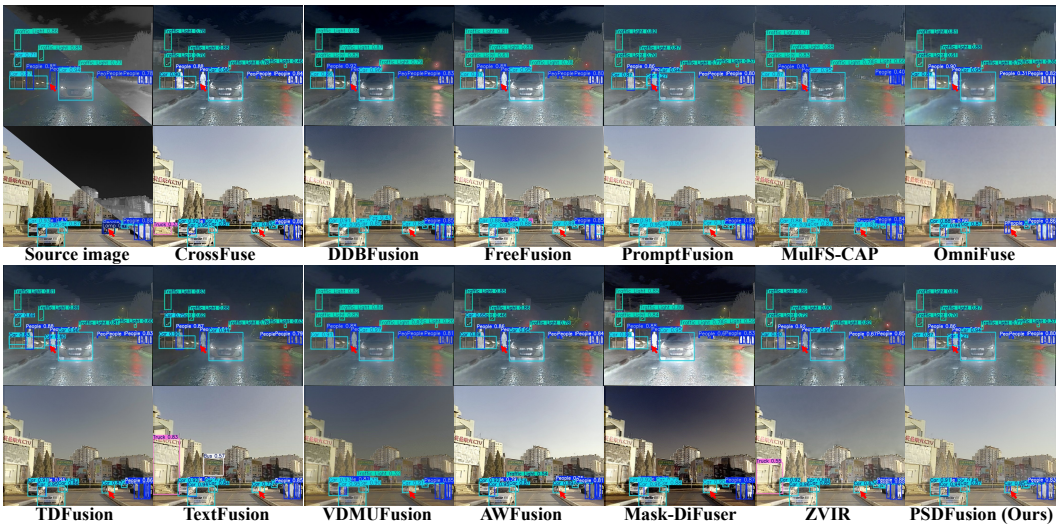

Figure 17: Qualitative comparison of object detection performance between the proposed PSDFusion and 12 state-of-the-art methods on M3FD.