

P²Fusion: Prompt-based Progressive Infrared-Visible Image Fusion via Dual-Prior Distillation

Yi Shi^{1†}, Huichao Xie^{1†}, Yuqing Wang¹, Mingyu Wang¹, Kaihui Yang¹, Yu Liu², Ruitao Lu³, Lizhe Li¹, Junwei Han^{1,4*}, and Dingwen Zhang^{1*}

¹ Northwestern Polytechnical University, Xi'an, China

² Hefei University of Technology, Hefei, China

³ Rocket Force University of Engineering, Xi'an, China

⁴ Chongqing University of Posts and Telecommunications, Chongqing, China

Abstract. Infrared-visible image fusion (IVIF) is pivotal for multimodal perception, yet reconciling the inherent information disparity between thermal and textural features remains a fundamental challenge. Existing prior-guided methods often rely on static constraints that induce optimization conflicts or utilize extrinsic semantic priors from large-scale foundation models (e.g., CLIP/DINO), which frequently fail to exploit the intrinsic modality characteristics essential for high-fidelity fusion. To address these issues, we propose P²Fusion, a prior-guided distillation-based framework that reformulates IVIF via dual intrinsic prompts. Instead of imposing hard-coded penalties, we distill image-intrinsic priors, thermal saliency and spatial quality—into learnable, dynamic regulators. Specifically, a Teach-to-Fuse mechanism provides dual-granularity progressive guidance, coupled with a Gated Dynamic Expert Recalibration (GDER) module for decoupled feature refinement. This design enables the network to adaptively mediate modal competition through expert specialization. Extensive experiments demonstrate that P²Fusion achieves state-of-the-art performance across five mainstream datasets. Notably, our framework demonstrates consistent performance advantages in fusion quality, achieving state-of-the-art results in 14 out of 20 key evaluation metrics across 5 benchmarks. Furthermore, it effectively contributes to the robustness of downstream perception, such as +3.2% mAP on MSRS, +0.5% mAP on M3FD and +0.9% mAP on DroneVehicle for object detection. Our code will be available at <https://github.com/YiShi99/P2Fusion>.

Keywords: Infrared-visible image fusion · Prompt-based learning · Knowledge distillation · Multimodal perception

¹ [†] These authors contributed equally to this work.

² ^{*} Corresponding authors: Junwei Han (jhan@nwpu.edu.cn), Dingwen Zhang (zdw2006yyy@nwpu.edu.cn).

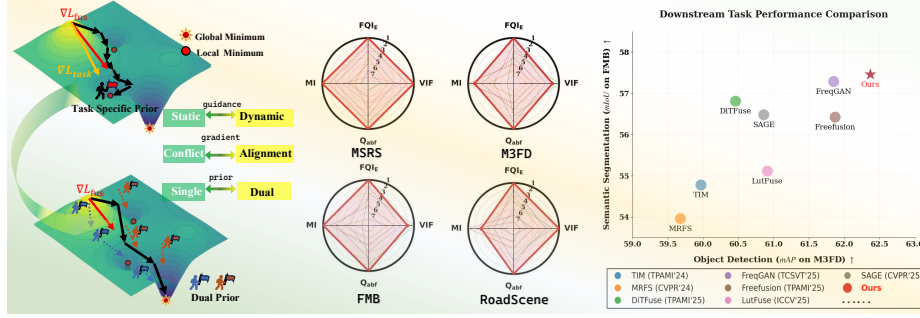


Fig. 1: Motivation and Performance Overview. (a) Prior integration paradigm (Left): Existing methods inject priors as static constraints, often causing optimization conflicts. Our P²Fusion provides dynamic, prompt-based guidance via image-intrinsic priors to mitigate gradient interference. (b) Fusion quality (Middle): Radar charts across four datasets (MSRS, M3FD, FMB, RoadScene) demonstrate our superior visual fidelity. (c) Downstream perception (Right): P²Fusion consistently enhances object detection and semantic segmentation performance on FMB and M3FD benchmarks.

1 Introduction

Multi-modal image fusion aims to integrate complementary information from diverse sensors to produce a comprehensive representation with rich details [22, 56]. This paradigm is indispensable in critical applications such as autonomous driving [4, 6, 8], target re-identification [3, 10], and medical imaging [62]. Among these, Infrared–Visible Image Fusion (IVIF) has emerged as a predominant research direction, leveraging the inherent synergy between thermal radiation and textural details [18, 30]. Since IVIF is essentially an unsupervised generative task due to the lack of paired ground-truth images, its performance heavily relies on the heuristic design of network architectures and loss functions [60, 65]. Without explicit labels, existing methods often struggle to balance detail fidelity and structural integrity, leading to the emergence of prior-guided fusion methods [24, 31, 43, 49]. By incorporating modality-specific cues, these methods provide indirect supervision to endow networks with scene-adaptive capabilities.

However, we observe that existing prior-guided paradigms face fundamental limitations across three mainstream technical routes. Firstly, methods based on explicit modality-specific hard-constraints [21, 31, 43] often treat segmentation masks or saliency maps as static regularization terms. This forces the network to prioritize prior distribution fitting over pixel-level feature integration, compelling the model to sacrifice fine-grained textures to satisfy rigid prior boundaries. Secondly, the route of joint optimization with downstream tasks [12, 18, 24, 49] integrates object detection or segmentation as auxiliary tasks. Nevertheless, these frameworks frequently introduce multi-objective optimization conflicts, where the drive for task-level accuracy compromises background detail preservation, creating a persistent trade-off between visual aesthetics and task performance. Thirdly, the recent trend of extrinsic semantic guidance [50, 56] leverages foun-

dation models (e.g., CLIP or DINO) as soft guidance. While mitigating the "hard-constraint" issue, it introduces extrinsic information that deviates from the modal essence. Since image fusion is fundamentally a pixel-level task, the abstract high-level representations lead to a granularity mismatch between semantics and pixels, preventing precise guidance for low-level textural reconstruction.

Based on these insights, we propose a novel Teach-to-Fuse (T2F) paradigm (Fig. 1) that returns to the essence of image data. Following this paradigm, we develop P²Fusion, a network built on the core philosophy that priors should be endogenous and dynamic. Rather than introducing extrinsic or rigid constraints, we focus on maximizing the mining of intrinsic attributes from the infrared and visible images themselves. We move away from seeking a "perfect" static prior and instead build an adaptive prompt learning framework. In this work, we select infrared thermal saliency and visible spatial quality as two complementary intrinsic priors. These two priors are chosen because they inherently capture distinctive modality traits and intuitively align with human visual perception, as supported by established literature [21] [43]. Crucially, our framework serves to instantiate a prior-guided prompt-based philosophy rather than chasing the "best" theoretical teachers. It is prior-agnostic and can seamlessly extend to other task-related auxiliary cues (e.g., depth, semantics), thereby releasing optimization freedom and achieving dynamic scene adaptation.

To further optimize fusion performance and strengthen cross-modal interactions, we design a fusion architecture featuring a Gated Dynamic Expert Recalibration (GDER) module, which functions as a self-correcting feedback mechanism. Unlike traditional Mixture-of-Experts (MoE) used for capacity expansion, our GDER module is tasked with feature decoupling and correction. Within this module, we synergize a prior-guided Modality Specialist with a global-aware Attention Specialist to perform modality-specific feedback refinement. By leveraging a reliability-based gating mechanism to adaptively recalibrate the fusion strategy, we effectively decouple conflicting features and rectify biased signals, thereby ensuring a robust and perceptually consistent fusion representation.

The main contributions of this work are summarized as follows:

- We propose the **"Teach-to-Fuse"** paradigm for multi-modal image fusion. By systematically addressing the pitfalls of hard-constrained priors, multi-objective conflicts, and granularity mismatch, we transform static penalties into process-oriented guidance that maximizes the mining of intrinsic modality attributes via dynamic prompts.
- We develop the **Gated Dynamic Expert Recalibration (GDER)** module to optimize feature refinement and cross-modal interactions. This MoE-based self-correction mechanism synergizes a prior-guided Modality Specialist with a global-aware Attention Specialist, enabling the network to adaptively decouple conflicting features and rectify biased signals through modality-specific feedback.
- **Extensive Benchmarking and Perceptual Analysis.** We conduct a comprehensive evaluation of the proposed framework across five public datasets and over ten benchmark tasks. The results consistently demonstrate the su-

periority of our Teach-to-Fuse paradigm in terms of high-fidelity visual fusion, robust cross-domain generalization, and significant performance gains in downstream perception tasks.

2 Related Work

2.1 Deep Learning Paradigms in IVIF

Deep learning has become the prevailing paradigm in IVIF, outperforming conventional methodologies such as sparse representation and multiscale decomposition [2, 34, 35, 54]. Early deep fusion frameworks primarily explored diverse architectures to improve feature representation, including Auto-Encoders (AEs) [13, 14, 25], CNNs [19, 55, 63], Generative Adversarial Networks (GANs) [17, 32, 53], and more recently, Transformers [30, 44, 62] and Diffusion Models [40, 64].

These advanced architectures have demonstrated notable advantages in synthesizing complementary information through attention mechanisms [45, 61], sophisticated feature interactions [15, 37, 65], and disentangled representation learning [5, 60]. However, most of these purely data-driven approaches rely on manually designed loss functions and heuristic architectural constraints to achieve fusion objectives [28, 29]. More comprehensive reviews and comparisons of IVIF methods can be found in [23].

2.2 Prior Knowledge Integration in Image Fusion

Prior-guided fusion has emerged as a pivotal paradigm in IVIF, substantially advancing both perceptual quality and downstream task performance [1, 20, 24, 56]. A dominant trend exploits high-level semantics from large-scale foundation models. Representative works include SAGE [49], which harnesses Segment Anything Model (SAM) [11] priors via semantic-persistent attention and dual-level distillation, and CMFS [7, 50], which employs CLIP [36] guidance for cross-modal interaction and frequency-spatial collaboration to suppress noise.

Complementary to foundation models, another line of research integrates priors from specialized vision tasks. STDFusionNet [31], for example, leverages saliency detection masks for spatial guidance and selective feature extraction. Others utilize semantic guidance from pre-trained segmentation models [22, 39, 42, 50] or employ object detection meta-features to mitigate feature misalignment [18, 61]. Beyond task-level semantics, alternative prior formulations are embedded within the network architecture. These include incorporating low-rank representation priors to guide feature decomposition [15], exploiting visible image quality priors [43], and exploring modality-specific priors within hybrid CNN-Transformer frameworks [58].

In summary, existing prior-guided IVIF methods often rely on rigid or extrinsic priors, which may cause optimization conflicts or granularity mismatch, leaving room for more adaptive, modality-intrinsic prior modeling.

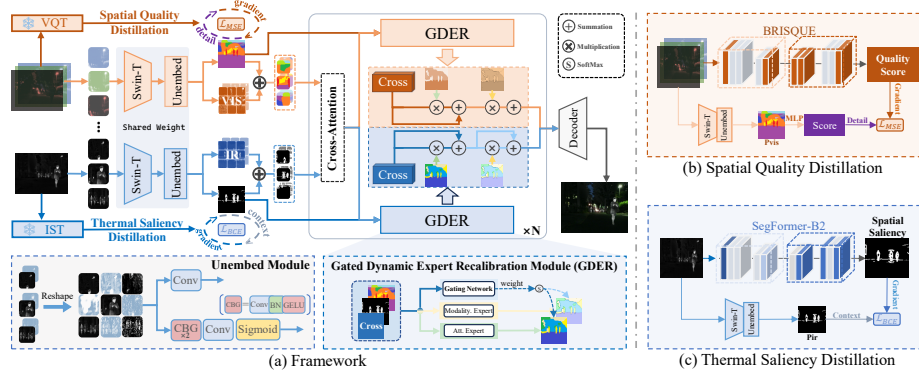


Fig. 2: Illustration of the proposed framework. (a) Overall pipeline: prior-guided visible and infrared features are fused via cross-attention, where the proposed GDER module adaptively refines modality-specific and interaction features for reconstruction. (b) Spatial Quality Distillation: a quality assessor provides global quality guidance to supervise spatial-detail learning. (c) Thermal Saliency Distillation: a saliency teacher offers thermal saliency priors to enhance target-aware fusion.

3 Methodology

In this section, we elucidate the proposed P²Fusion framework, which instantiates the "Teach-to-Fuse" paradigm. The architecture, as illustrated in Fig. 2, consists of two core stages: (1) **Dual-Teacher Prompt Distillation**, which transforms static priors into learnable, process-oriented guidance signals; and (2) **Adaptive Fusion with Gated Dynamic Expert Recalibration (GDER)**, which achieves robust feature integration and self-correcting refinement via decoupled specialized experts.

3.1 Dual-Teacher Prompt Distillation

In conventional prior-guided methods, knowledge (e.g., saliency maps) is typically injected as rigid spatial constraints, which limits the optimization flexibility of the network. We argue that prior knowledge should participate in the fusion process as soft, dynamic prompts. We select two complementary intrinsic priors as a validation instance:

- Infrared Saliency Prior (T_{ir}): Derived from a SegFormer-B2 [51] pre-trained on the representative MSRS [41] and FMB [22] datasets, this prior serves as the Infrared-Saliency Teacher (IST) to localize high-thermal-radiation targets (e.g., pedestrians and vehicles) in the infrared modality.
- Visible Quality Prior (T_{vis}): Generated by the no-reference Visible-Quality Teacher (VQT) based on the BRIQUE [33] evaluator, this prior assesses textural reliability spatially, such as underexposed or overexposed areas.

For the visible branch, the VQT generates a global quality score $T_{vis} \in \mathbb{R}$. We project the high-dimensional prompt P_{vis} into a scalar P'_{vis} to match the teacher’s output. For the infrared branch, the IST provides a saliency map $T_{ir} \in \mathbb{R}^{H \times W}$. The final enriched features $F = \{F_{vis}, F_{ir}\}$ are obtained via element-wise summation: $F_m = f_m + \text{Embed}(P_m)$, where $m \in \{vis, ir\}$.

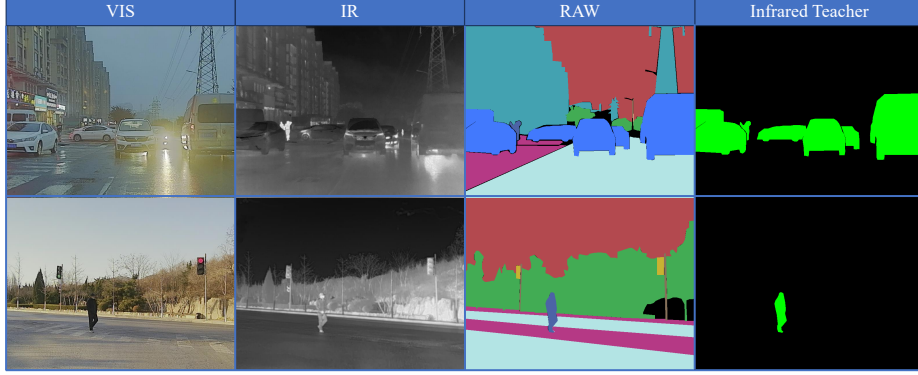


Fig. 3: Illustration of the difference between the original dataset and that used for training our infrared thermal saliency teacher: pedestrians and vehicles are defined as foreground, while all others are treated as background.

Specifically, for the infrared branch, semantic predictions are post-processed into binary saliency masks to incorporate domain priors, where pedestrians and vehicles are treated as foreground and all remaining categories as background as shown in Fig. 3. The resulting saliency map is then downsampled via bilinear interpolation to match the spatial resolution of the infrared prompt P_{ir} . For the visible branch, given an image-level score $s \in [0, 100]$, we normalize it into a continuous quality prior by $T_{vis} = (100 - s)/100$, where higher values correspond to stronger textural reliability.

3.2 Adaptive Fusion with GDER Refinement

The feature spaces of infrared and visible modalities are inherently heterogeneous—one being sensitive to thermal contours and the other rich in textural details. To mitigate modality confusion, we maintain dual-branch processing and employ cross-attention (CA) [27, 46] mechanism for feature modulation:

$$F'_{ir} = \text{CA}(F_{ir}, F_{vis}, F_{vis}), \quad F'_{vis} = \text{CA}(F_{vis}, F_{ir}, F_{ir}) \quad (1)$$

where $\text{CA}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$. This stage yields cross-modal interactions $F' = \{F'_{vis}, F'_{ir}\}$. To further refine the fused representations and capture modal-specific nuances, we propose the Gated Dynamic Expert Recalibration

(GDER) module. Unlike Mixture-of-Experts (MoE) architectures designed primarily for model capacity expansion, our GDER module functions as a decoupled feature recalibrator with explicit functional specialization, aiming to adaptively aggregate the most informative cues from heterogeneous feature spaces:

1. **Prior-Responsive Modality Expert (E_{mod}):** This expert explicitly consumes the dynamic prompts to reinforce modality-specific salient information (e.g., thermal targets or high-contrast textures). It ensures that the network maintains a high-fidelity focus on regions indicated by the learned priors, facilitating effective knowledge injection.
2. **Prior-Agnostic Attention Expert (E_{att}):** Operating independently of the prompts, this expert utilizes a global attention mechanism (CBAM) to mine intrinsic long-range dependencies and fine-grained structural details. It acts as a contextual compensator to ensure the integrity of textures that may fall outside the prior’s primary focus.

The GDER module acts as an adaptive integration arbiter. A gating network $G(\cdot)$ computes dynamic weights $w = \text{softmax}(G(F', P))$ based on the mutual agreement between local feature characteristics and global prompt signals. The refined feature is formulated as:

$$F'' = F' + w_1 \cdot E_{mod}(F') + w_2 \cdot E_{att}(F') \quad (2)$$

By synergizing the outputs of these decoupled specialists, the GDER module adaptively recalibrates bimodal information flow. The Prior-Responsive Expert anchors features on saliency-guided regions, while the Prior-Agnostic Expert captures the underlying structural dependencies that might be overlooked by prompts. This dual-expert design allows the network to transcend the limitations of any single representation, achieving robust feature recalibration that balances salient foreground targets with rich background textures.

In this configuration, the gating network acts as a dynamic weight allocator that optimizes the integration of modality-specific and context-aware features. When both specialists provide congruent information, they work in synergy to enhance feature representational power; when their focuses diverge, the gating mechanism adaptively balances their contributions to ensure that the fused feature remains both semantically meaningful and structurally complete. This mechanism achieves adaptive feature refinement through multi-perspective complementarity, rather than rigid constraint. After N iterative refinements, the refined dual-modality features are concatenated and fed into the reconstruction decoder to generate the final fused image I_{fused} .

3.3 Optimization Objectives

We replace traditional hard-coded prior losses with learnable prompt guidance. The total loss $\mathcal{L}_{total}$ comprises perceptual constraints and distillation terms.

Perceptual Constraints: The Gradient Loss ensures edge preservation: $\mathcal{L}_{grad} = \|\nabla I_{fused} - \max(|\nabla I_{ir}|, |\nabla I_{vis}|)\|_1$. The Intensity Loss balances thermal

and texture information: $\mathcal{L}_{int} = \|I_{fused} - \max(I_{ir}, I_{vis})\|_1$. We further introduce the SSIM Loss to maintain structural consistency:

$$\mathcal{L}_{ssim} = 1 - (\omega_{ir} \text{SSIM}(I_{ir}, I_{fused}) + \omega_{vis} \text{SSIM}(I_{vis}, I_{fused})) \quad (3)$$

where weights ω are adaptively computed based on the mean gradient magnitude of each modality.

Prompt Distillation Loss: The visible prompts are optimized via MSE: $\mathcal{L}_{vis}^{distill} = \|P'_{vis} - T_{vis}\|_2^2$. The infrared prompts are optimized via Binary Cross-Entropy (BCE) to ensure spatial saliency alignment:

$$\mathcal{L}_{ir}^{distill} = -(T_{ir} \odot \log P_{ir}^{map} + (1 - T_{ir}) \odot \log(1 - P_{ir}^{map})) \quad (4)$$

The overall optimization objective is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{int} + \lambda_2 \mathcal{L}_{ssim} + \lambda_3 \mathcal{L}_{grad} + \lambda_4 \mathcal{L}_{ir}^{distill} + \lambda_5 \mathcal{L}_{vis}^{distill} \quad (5)$$

4 Experiments

4.1 Experiments Settings

Datasets and metrics. We evaluate P²Fusion on five public benchmarks: MSRS [41], M3FD [18], FMB [22], RoadScene [52] and DroneVehicle [38]. Generalization is assessed on their official test splits, covering fusion quality (all), object detection (MSRS, M3FD, DroneVehicle), and semantic segmentation (MSRS, FMB). Fusion quality is measured using four standard metrics, while downstream evaluations follow the respective standard protocols.

Implementations. Our framework is implemented in PyTorch on four NVIDIA RTX 3090 GPUs. Adam optimizes parameters for 100 epochs (batch size 12, initial lr 1×10^{-4}), decayed by 0.5 at iterations [3000, 6000, 9000, 12000, 15000]. The optimal weights for the total loss (λ_1 – λ_5) are (8,25,20,0.5,5).

Comparisons. We compare with 12 SOTA fusion networks (SAGE [49], DiTFuse [16], MRFS [59], Freefusion [60], FreqGAN [48], LutFuse [57], LRRNet [15], DDFM [64], TarDal [18], ReCoNet [9], TIM [26]) to validate our method’s effectiveness and generalization ability.

4.2 Infrared and visible image fusion

We conduct IVIF experiments on MSRS (361 pairs), M3FD (300), FMB (280), RoadScene (221) and DroneVehicle (8980). Training uses combined MSRS (1083 pairs) and FMB (1220) training sets, whose segmentation labels train the infrared saliency teacher as mentioned in 3.1.

Qualitative Comparisons. Qualitative results for selected pairs are in Fig. 4 (first row) and Fig. 5. On the MSRS dataset, SAGE, DiTFuse, FreqGAN, LRRNet, FreeFusion, and TIM exhibit an over-reliance on visible textures at the expense of infrared saliency, resulting in attenuated structural details in low-light

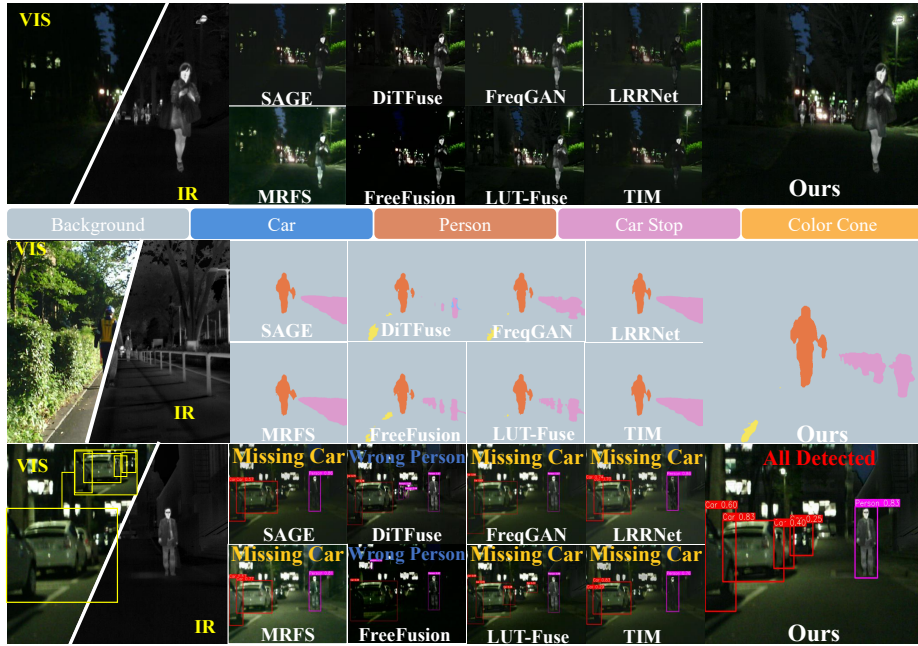


Fig. 4: Comprehensive qualitative comparison on the MSRS dataset. Row 1 shows fused images; Row 2 and Row 3 show the corresponding semantic segmentation and object detection results, respectively. Our method consistently achieves superior performance across image fusion, semantic segmentation, and object detection tasks.

regions. Conversely, MRFS yields over-smoothed textures with suboptimal local contrast, while LUT-Fuse introduces noticeable artifacts around moving targets despite better bimodal preservation. In contrast, our method effectively recovers background details (e.g., tree branches and shadowed building textures) while maintaining sharp pedestrian contours and high fidelity.

In the high-contrast M3FD scene (Fig. 5, first row), competing methods suffer from modality collapse, making the background text indistinct. Our method maintains a stable equilibrium, yielding clearer structures and superior text legibility.

In the smoke-occluded FMB scene (Fig. 5, second row), most competitors (SAGE, MRFS, FreqGAN, LUT-Fuse, LRRNet, TIM) under-exploit infrared cues, leaving details corrupted by smoke. Meanwhile, DiTFuse and FreeFusion over-rely on the infrared modality, inducing an unnatural night-like shift and losing visible textures like the left tree. Our method ensures an optimal balance, simultaneously preserving infrared contours and fine visible details.

In the over-exposed RoadScene scene (Fig. 5, third row), SAGE, MRFS, FreeFusion, FreqGAN, and LUT-Fuse under-utilize infrared cues, resulting in over-exposed outputs that miss critical contours (e.g., the pedestrian near the tree trunk). In contrast, DiTFuse, LRRNet, and TIM fail to maintain modality

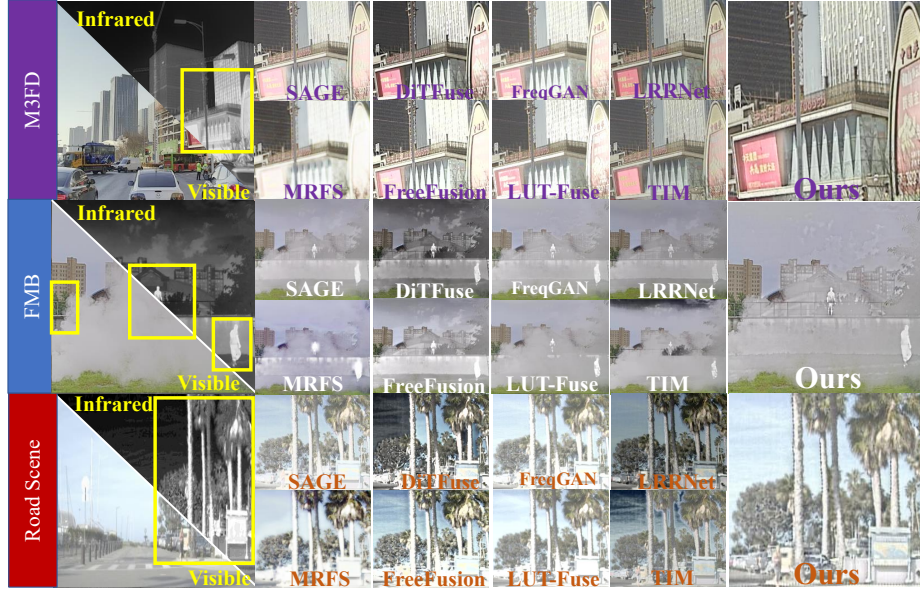


Fig. 5: Qualitative fusion comparison. Representative high-contrast scenes from M3FD, smoke-occluded scenes from FMB, and over-exposed scenes from RoadScene are selected. Our method consistently preserves infrared salient targets and visible textures across degradations, achieving the best modality balance.

Table 1: Quantitative comparison on the MSRS, M3FD, FMB, and RoadScene datasets using four key evaluation metrics. (1st: Red, 2nd: Light Blue)

Datasets	MSRS				M3FD				FMB				RoadScene			
Methods	$FQI_E \uparrow$	$VIF \uparrow$	$Q_{abf} \uparrow$	$MI \uparrow$	$FQI_E \uparrow$	$VIF \uparrow$	$Q_{abf} \uparrow$	$MI \uparrow$	$FQI_E \uparrow$	$VIF \uparrow$	$Q_{abf} \uparrow$	$MI \uparrow$	$FQI_E \uparrow$	$VIF \uparrow$	$Q_{abf} \uparrow$	$MI \uparrow$
DiTFuse(TPAMI'25)	0.642	0.281	0.363	2.146	0.389	0.188	0.274	2.514	0.536	0.215	0.363	2.620	0.445	0.320	0.426	2.938
FreqGAN(TCSVT'25)	0.679	0.285	0.492	2.871	0.445	0.282	0.399	3.272	0.518	0.293	0.437	3.434	0.442	0.279	0.444	2.967
FreeFusion(TPAMI'25)	0.590	0.281	0.430	1.984	0.621	0.473	0.539	2.712	0.696	0.501	0.587	3.004	0.403	0.362	0.403	2.692
LutFuse(ICCV'25)	0.783	0.424	0.610	3.625	0.466	0.319	0.480	3.987	0.575	0.319	0.527	3.897	0.328	0.265	0.381	3.644
SAGE(CVPR'25)	0.763	0.335	0.535	3.217	0.659	0.363	0.575	3.115	0.763	0.384	0.634	3.429	0.379	0.237	0.334	2.919
TIM(TPAMI'24)	0.655	0.277	0.469	3.025	0.533	0.325	0.524	3.533	0.679	0.125	0.581	3.534	0.297	0.181	0.354	3.611
MRFS(CVPR'24)	0.625	0.340	0.484	3.068	0.287	0.194	0.249	2.965	0.477	0.161	0.388	3.087	0.268	0.159	0.271	2.788
LRRNet(TPAMI'23)	0.599	0.243	0.420	2.938	0.530	0.283	0.483	2.823	0.649	0.112	0.539	3.024	0.277	0.133	0.326	2.791
DDFM(ICCV'23)	0.569	0.363	0.469	2.666	0.522	0.311	0.456	2.851	0.566	0.214	0.504	3.144	0.485	0.372	0.471	2.940
TarDal(CVPR'22)	0.495	0.357	0.428	2.646	0.425	0.317	0.402	3.182	0.443	0.304	0.406	3.419	0.395	0.328	0.392	3.404
ReCoNet(ECCV'22)	0.720	0.325	0.500	3.115	0.570	0.291	0.481	3.095	0.692	0.168	0.546	3.288	0.422	0.245	0.334	3.139
Ours	0.850	0.445	0.682	3.905	0.743	0.429	0.635	3.819	0.826	0.448	0.677	4.030	0.617	0.388	0.557	3.021

balance, producing unnaturally dark skies and losing details on billboards. Our method achieves a more stable dynamic balance, recovering the global appearance while preserving rich textures and fine details.

Quantitative Comparisons. Tables 1 report the quantitative results. Across all datasets, our method achieves SOTA on 12 out of 16 metrics and ranks second on 3 others, highlighting its overall superiority and strong cross-dataset generalization.

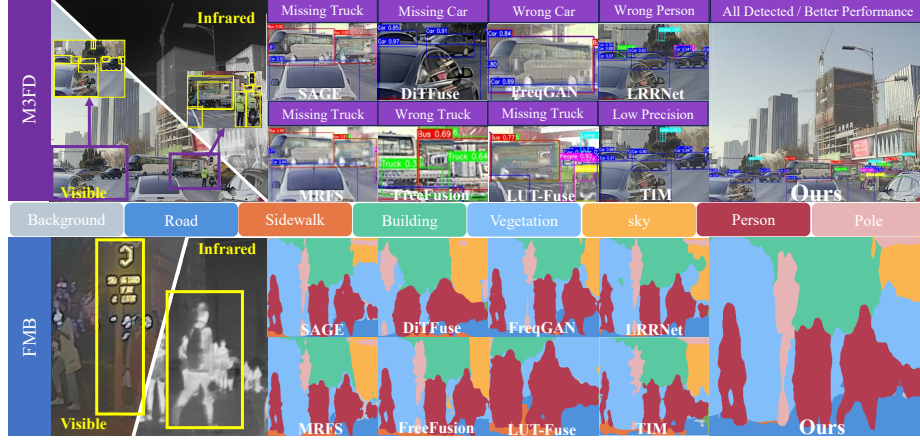


Fig. 6: Qualitative evaluation on downstream tasks. We visualize object detection results on M3FD (Row 1) and semantic segmentation results on FMB (Row 2). Our framework demonstrates superior performance in downstream tasks.

Table 2: Quantitative results of downstream tasks: Object Detection on M3FD and Semantic Segmentation on FMB. (1st: Red, 2nd: Light Blue)

Datasets	M3FD(Object Detection)							FMB(Semantic Segmentation)						
Methods	Person	Car	Bus	Lamp	Moto.	Truck	mAP↑	T.Sign	Person	Car	Truck	Moto.	Pole	mPA↑ mIoU↑
DiTFuse(TPAMI'25)	0.514	0.684	0.742	0.497	0.525	0.663	60.46	73.146	63.568	81.763	33.809	38.908	43.214	63.340 56.808
FreqGAN(TCSVT'25)	0.544	0.700	0.766	0.510	0.519	0.673	61.85	73.558	66.301	82.225	31.691	38.425	44.730	64.471 57.284
FreeFusion(TPAMI'25)	0.527	0.702	0.770	0.526	0.512	0.674	61.87	71.751	64.320	80.160	36.174	33.359	41.968	63.469 56.421
LutFuse(ICCV'25)	0.530	0.691	0.763	0.500	0.501	0.670	60.91	72.434	65.192	80.893	16.908	38.130	42.354	62.221 55.113
SAGE(CVPR'25)	0.531	0.707	0.735	0.483	0.539	0.658	60.86	72.390	65.286	81.387	31.706	35.223	41.864	63.798 56.479
TIM(TPAMI'24)	0.496	0.686	0.743	0.503	0.506	0.665	59.97	72.216	62.701	81.081	26.209	30.750	41.752	61.942 54.776
MRFS(CVPR'24)	0.508	0.687	0.760	0.482	0.483	0.657	59.68	67.617	63.632	79.955	28.056	33.916	40.510	60.926 53.959
LRRNet(TPAMI'23)	0.519	0.693	0.756	0.514	0.499	0.659	60.67	70.693	64.877	81.654	34.215	31.684	41.634	63.853 56.382
DDFM(ICCV'23)	0.526	0.706	0.749	0.483	0.558	0.662	61.44	71.478	65.847	81.797	32.143	33.463	42.865	64.466 57.001
TarDal(CVPR'22)	0.544	0.693	0.755	0.512	0.495	0.672	61.22	72.249	66.804	80.383	15.760	34.881	41.201	62.462 55.193
ReCoNet(ECCV'22)	0.525	0.704	0.755	0.529	0.500	0.679	61.53	71.570	64.709	81.360	28.587	32.394	41.458	63.631 56.113
Ours	0.534	0.704	0.773	0.521	0.521	0.690	62.37	74.153	65.670	82.270	38.270	37.024	43.913	64.672 57.456

4.3 IVIF for Downstream Tasks

For object detection, we train YOLOv7s [47] on MSRS-detection (80 pairs) and M3FD (4200 pairs) with an 8:2 split. Tables 2 (left) and 3 (left) show that our method achieves the highest mAP₅₀₋₉₅, outperforming the previous SOTA by +3.2% on MSRS and +0.5% on M3FD. Visualizations in Figs. 4 (third row) and 6 confirm that our fused images successfully detect all scene targets, whereas other methods suffer from false positives or missed detections.

For semantic segmentation, SegFormer-B5 [51] is adopted on MSRS (1083/361 pairs) and FMB (1220/280). As reported in Tables 2 (right) and 3 (right), our framework delivers the best performance with consistent gains. Qualitative results (Figs. 4, 6, second rows) demonstrate more coherent masks with better structural preservation and reduced fragmentation.

Table 3: Quantitative comparison on the MSRS datasets for object detection and segmentation. (1st: Red, 2nd: Light Blue)

Tasks	Object Detection (mAP_{50} $mAP_{50 \rightarrow 95}$)						Semantic Segmentation					
Methods	Person		Car		All		Car	Person	Curve	Stop	Cone	mPA $\uparrow$ mIoU $\uparrow$
DiTFuse(TPAMI'25)	0.993	0.717	0.838	0.715	0.915	0.716	88.073	68.929	40.988	53.430	66.992	70.555 62.390
FreqGAN(TCSVT'25)	0.995	0.731	0.906	0.645	0.951	0.688	87.681	69.427	33.235	62.014	66.369	70.149 62.100
FreeFusion(TPAMI'25)	0.991	0.723	0.763	0.385	0.877	0.554	84.702	66.407	36.520	64.899	64.334	68.776 61.120
LutFuse(ICCV'25)	0.995	0.727	0.937	0.628	0.966	0.677	87.848	70.602	44.679	59.176	67.835	71.558 63.471
SAGE(CVPR'25)	0.991	0.700	0.609	0.476	0.800	0.588	91.265	66.107	27.288	29.242	56.066	63.055 56.228
TIM(TPAMI'24)	0.986	0.603	0.920	0.643	0.953	0.623	89.800	62.801	31.877	32.128	56.872	61.471 55.267
MRFS(CVPR'24)	0.991	0.738	0.920	0.614	0.955	0.676	91.222	65.075	26.845	29.521	56.422	62.333 55.073
LRRNet(TPAMI'23)	0.994	0.757	0.962	0.606	0.978	0.681	91.705	67.741	30.996	31.368	57.681	62.450 56.567
DDFM(ICCV'23)	0.986	0.697	0.747	0.585	0.866	0.641	91.911	68.686	31.325	28.690	56.075	63.932 56.875
TarDal(CVPR'22)	0.974	0.716	0.858	0.669	0.916	0.692	91.304	68.907	28.039	28.831	55.925	62.784 56.534
ReCoNet(ECCV'22)	0.994	0.712	0.866	0.573	0.931	0.643	91.750	66.247	30.113	32.249	56.779	62.844 56.928
VIS	0.851	0.477	0.862	0.626	0.856	0.552	91.667	57.770	25.817	29.342	59.742	60.604 54.559
IR	0.979	0.613	0.962	0.672	0.971	0.643	87.155	68.749	25.281	22.444	45.700	54.172 48.778
Ours	0.995	0.776	0.995	0.720	0.995	0.748	88.520	70.030	40.903	59.800	69.494	71.480 63.793

4.4 Ablation Study

Study on Prior Knowledge Integration. Fig. 7(a) further demonstrates that removing either prior consistently degrades performance across all metrics. Specifically, excluding the visible prior (W/O vis) causes a sharp drop in VIF , highlighting its importance in preserving fine-grained textures and structural fidelity. In contrast, the full ablation of both infrared and visible priors (W/O ir+vis) yields the worst performance, indicating that prior guidance is essential for balanced fusion quality.

Fig. 8 demonstrates the effectiveness of bimodal priors in orthogonal gating. The full model (a) achieves structured, spatially selective gating. Removing P_{vis} (b) causes the gating to collapse into a near-uniform distribution, confirming its role in providing localized guidance. Conversely, removing P_{ir} (c) results in over-smoothed IR responses and a loss of cross-modal enhancement, highlighting its importance for discriminative performance.

Study on Gated Dynamic Expert Recalibration. Fig. 7(b) compares our GDER module against intuitive alternatives, such as applying patch-level rather than global-level loss to P_{vis} ("Distribution"). Empirical results confirm the superiority of our approach. Fig. 7(c) validates the GDER teacher framework, where the "W/O both" configuration shows the largest performance drop, underscoring the architecture's importance. Additionally, a Pearson correlation of $r = 0.46$ between the Modality and Attention experts demonstrates effective functional decoupling. Our full model achieves the optimal score across all metrics, confirming that our dual-teacher strategy provides comprehensive, non-redundant guidance.

Fig. 9 visualizes expert activation maps and prior-guided prompts. Complementarity within one modality and distinct responses across modalities validate our decoupled specialization over naive fusion. In overexposed scenes, the attention expert captures structural details while the modality expert isolates thermal targets, proving they are synergistically coupled yet representationally

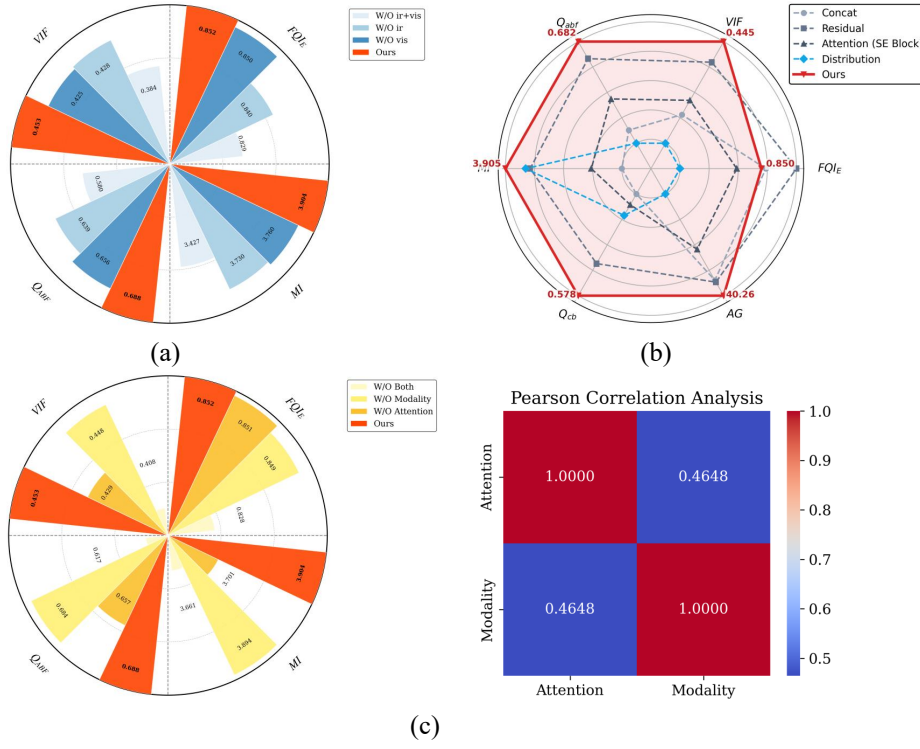


Fig. 7: Ablation study of the proposed components. All metrics are normalized relative to the full model (Ours). (a) **Impact of image priors:** Dual priors show clear synergy in maintaining visual fidelity. (b) **Structural alternatives to GDER:** Comparison with simpler architectural designs. (c) **Analysis of GDER teachers:** (Left) Performance gains from Modality and Attention teachers. (Right) Correlation analysis verifying the functional decoupling and non-redundancy between experts.

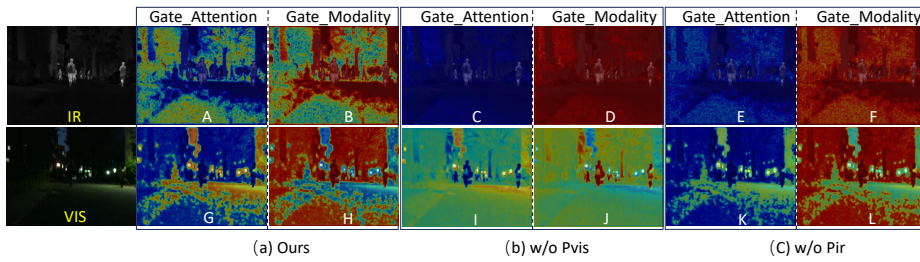
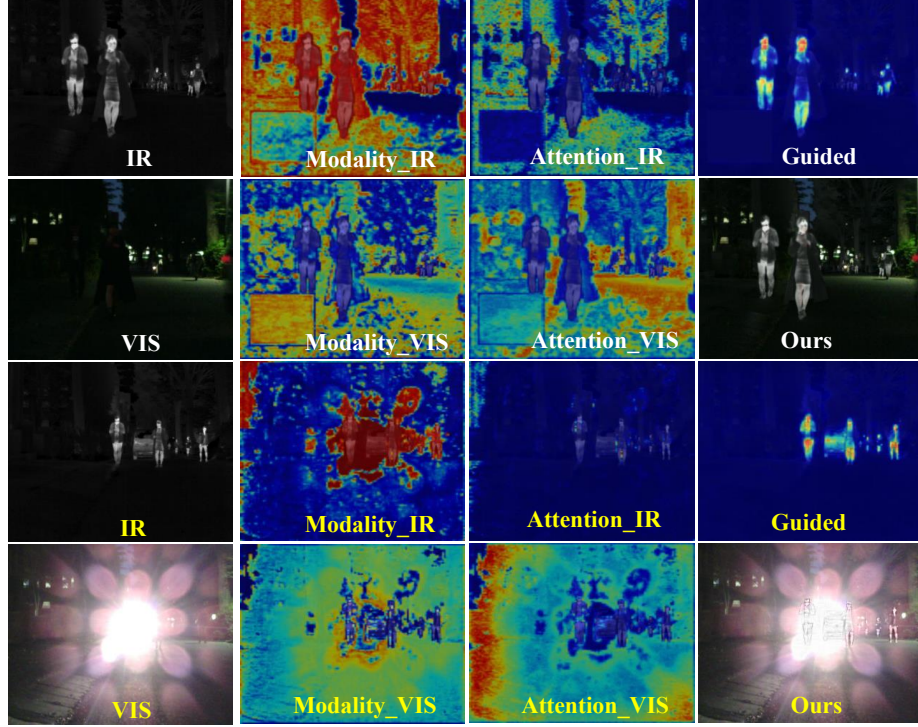


Fig. 8: Visual analysis of the gating mechanism. Visualization of cross-modal gating allocations, where ablating P_{vis} or P_{ir} disrupts spatial discrimination and leads to uniform or blurred feature distributions.

decoupled. Guided by high-fidelity spatial anchors, our framework ensures robust cross-modal integration.

Table 4: Quantitative results of OOD Verification on the DroneVehicle, including image fusion and downstream object detection results. (1st: Red, 2nd: Light Blue)

Tasks	Image Fusion Metrics				Object Detection				
Methods	$FQI_E \uparrow$	$VIF \uparrow$	$Q_{abf} \uparrow$	$MI \uparrow$	Car	Truck	Bus	Van	$mAP_{50 \rightarrow 95} \uparrow$
DiTFuse (TPAMI'25)	0.3929	0.1556	0.3134	2.1892	0.382	0.354	0.578	0.234	0.387
FreqGAN (TCSVT'25)	0.5340	0.2775	0.4643	2.6642	0.601	0.490	0.710	0.360	0.540
FreeFusion (TPAMI'25)	0.4893	0.2690	0.4339	2.4806	0.558	0.377	0.659	0.260	0.463
LutFuse (ICCV'25)	0.5203	0.2955	0.4984	3.8038	0.571	0.411	0.681	0.292	0.489
SAGE (CVPR'25)	0.4443	0.1791	0.3636	2.4162	0.582	0.447	0.694	0.324	0.512
TIM (TPAMI'24)	0.2728	0.2211	0.3182	2.6823	0.622	0.567	0.729	0.463	0.595
MRFS (CVPR'24)	0.3832	0.1790	0.3268	2.4254	0.562	0.383	0.676	0.263	0.471
LRRNet (TPAMI'23)	0.2490	0.1937	0.2848	2.1306	0.568	0.400	0.666	0.245	0.470
DDFM (ICCV'23)	0.4549	0.2515	0.3829	2.5912	0.597	0.487	0.705	0.363	0.538
TarDal (CVPR'22)	0.3992	0.2160	0.3576	2.7220	0.578	0.451	0.694	0.329	0.513
ReCoNet (ECCV'22)	0.4129	0.1982	0.3395	2.3772	0.581	0.435	0.701	0.289	0.501
Ours	0.6282	0.2655	0.5269	2.9998	0.632	0.576	0.740	0.469	0.604

**Fig. 9: Qualitative analysis of expert decoupling and feature visualization.** We present the activation maps for the Modality and Attention experts across infrared and visible branches. The distinct focus of each expert—targeting either thermal saliency or structural textures—validates the functional complementarity of our MoE-based decoupling strategy.

4.5 Robustness and Out-of-Distribution (OOD) Verification

Prior-guided IVIF frameworks are often sensitive to prior quality, risking performance degradation in degraded scenes. To verify the resilience of P^2 Fusion, we evaluate it under challenging scenarios like extreme overexposure and smoke occlusion (Figs. 5, 9). Our framework mitigates prior dependency via the Teach-to-Fuse paradigm and its error-correction capability, as evidenced by SOTA results on M3FD and RoadScene without domain-specific fine-tuning (Table 1).

To test adaptability, we performed rigorous out-of-distribution (OOD) verification on the **DroneVehicle** dataset. This benchmark introduces a severe domain shift with aerial perspectives and severe visibility degradations. As summarized in Table 4, P^2 Fusion consistently maintains superiority without any prior domain exposure during training. Specifically, in image fusion, our method achieves SOTA on two metrics and second place on another among five primary indicators. More importantly, in downstream object detection, P^2 Fusion achieves a comprehensive SOTA across all categories, outperforming the second-best competitor by 0.9% in mAP_{50-95} . By dynamically mediating between intrinsic prompts and raw features, P^2 Fusion exhibits exceptional self-adaptability in unseen, hostile environments.

5 Conclusion

In this paper, we rethink the role of prior knowledge in infrared-visible image fusion and propose P^2 Fusion, a novel framework that shifts the fusion paradigm from static prior constraints to dynamic intrinsic prompting. By introducing a Teach-to-Fuse mechanism, we successfully distill task-specific physical priors—thermal saliency and spatial quality—into learnable regulators that procedurally govern the fusion process. Central to our architecture is the Gated Dynamic Expert Recalibration (GDER) module, which adaptively mediates modal competition and rectifies feature conflicts through expert specialization.

Extensive evaluations across five benchmarks and various challenging scenarios—including overexposure, smoke occlusion, and rigorous out-of-distribution (OOD) verification demonstrate that P^2 Fusion not only achieves superior visual fidelity but also significantly enhances downstream perception. The remarkable generalization capability in unseen domains further underscores that our method effectively mitigates the inherent prior-dependency of traditional frameworks. Ultimately, this work provides a robust and versatile solution for multi-modal information integration, bridging the gap between high-level physical guidance and adaptive feature representation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62293543 and Grant 62322605.

References

1. Chen, J., Yang, L., Yu, W., Gong, W., Cai, Z., Ma, J.: Sdsfusion: A semantic-aware infrared and visible image fusion network for degraded scenes. *IEEE Transactions on Image Processing* **34**, 3139–3153 (2025). <https://doi.org/10.1109/TIP.2025.3571339>
2. Cvejic, N., Bull, D., Canagarajah, N.: Region-based multimodal image fusion using ica bases. *IEEE Sensors Journal* **7**(5), 743–751 (2007). <https://doi.org/10.1109/JSEN.2007.894926>
3. Feng, J., Wu, A., Zheng, W.S.: Shape-erased feature learning for visible-infrared person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22752–22761 (2023). <https://doi.org/10.1109/CVPR52729.2023.02179>
4. Gao, H., Cheng, B., Wang, J., Li, K., Zhao, J., Li, D.: Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics* **14**(9), 4224–4231 (2018). <https://doi.org/10.1109/TII.2018.2822828>
5. Gao, Y., Ma, S., Liu, J.: Dcdr-gan: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology* **33**(2), 549–561 (2023). <https://doi.org/10.1109/TCSVT.2022.3206807>
6. Gao, Y., Dai, Y., Li, H., Ye, W., Chen, J., Chen, D., Zhang, D., He, T., Zhang, G., Han, J.: Cosurfgs: 3d surface gaussian splatting with collaborative distributed learning for large-scale scene reconstruction. *International Journal of Computer Vision* **134**, 195 (2026). <https://doi.org/10.1007/s11263-025-02627-9>
7. Guilin Su, Yuqing Huang, C.Y., He, Z.: Cmfs: Clip-guided modality interaction for mitigating noise in multi-modal image fusion and segmentation. In: *Proceedings of the Thirty-four International Joint Conference on Artificial Intelligence* (2025)
8. Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T.: Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. pp. 5108–5115 (2017). <https://doi.org/10.1109/IRROS.2017.8206396>
9. Huang, Z., Liu, J., Fan, X., Liu, R., Zhong, W., Luo, Z.: Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In: *Computer Vision – ECCV 2022*. pp. 539–555. Springer Nature Switzerland, Cham (2022)
10. Ji, D., Wang, H., Hu, H., Gan, W., Wu, W., Yan, J.: Context-aware graph convolution network for target re-identification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. pp. 1646–1654 (2021). <https://doi.org/10.1609/aaai.v35i2.16257>
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. *arXiv:2304.02643* (2023)
12. Li, H., Qin, R., Zou, Z., He, D., Li, B., Dai, B., Zhang, D., Han, J.: Langsurf: Language-embedded surface gaussians for 3d scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (01), 1–13 (2026). <https://doi.org/10.1109/TPAMI.2026.3701287>
13. Li, H., Wu, X.J.: Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing* **28**(5), 2614–2623 (2019). <https://doi.org/10.1109/TIP.2018.2887342>

14. Li, H., Wu, X.J., Kittler, J.: Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion* **73**, 72–86 (2021). <https://doi.org/10.1016/j.inffus.2021.02.023>, <https://www.sciencedirect.com/science/article/pii/S1566253521000440>
15. Li, H., Xu, T., Wu, X.J., Lu, J., Kittler, J.: 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). <https://doi.org/10.1109/TPAMI.2023.3268209>
16. Li, J., Jiang, C., Jiang, J., Liang, P., Ma, J., Nie, L.: Towards unified semantic and controllable image fusion: A diffusion transformer approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–18 (2025). <https://doi.org/10.1109/TPAMI.2025.3642842>
17. Li, Q., Lu, L., Li, Z., Wu, W., Liu, Z., Jeon, G., Yang, X.: Coupled gan with relativistic discriminators for infrared and visible images fusion. *IEEE Sensors Journal* **21**(6), 7458–7467 (2021). <https://doi.org/10.1109/JSEN.2019.2921803>
18. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5792–5801 (2022). <https://doi.org/10.1109/CVPR52688.2022.00571>
19. Liu, J., Fan, X., Jiang, J., Liu, R., Luo, Z.: Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(1), 105–119 (2022). <https://doi.org/10.1109/TCSVT.2021.3056725>
20. Liu, J., Li, X., Wang, Z., Jiang, Z., Zhong, W., Fan, W., Xu, B.: Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica* **12**(3), 502–515 (2025). <https://doi.org/10.1109/JAS.2024.124878>
21. Liu, J., Lin, R., Wu, G., Liu, R., Luo, Z., Fan, X.: Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision* **132**(5), 1748–1775 (2023). <https://doi.org/10.1007/s11263-023-01952-1>, <https://doi.org/10.1007/s11263-023-01952-1>
22. Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., Luo, Z., Fan, X.: Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8081–8090 (2023). <https://doi.org/10.1109/ICCV51070.2023.00745>
23. Liu, J., Wu, G., Liu, Z., Wang, D., Jiang, Z., Ma, L., Zhong, W., Fan, X., Liu, R.: Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(4), 2349–2369 (2025). <https://doi.org/10.1109/TPAMI.2024.3521416>
24. Liu, J., Zhang, B., Mei, Q., Li, X., Zou, Y., Jiang, Z., Ma, L., Liu, R., Fan, X.: Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2226–2235 (2025). <https://doi.org/10.1109/CVPR52734.2025.00213>
25. Liu, R., Liu, Z., Liu, J., Fan, X.: Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In: *Proceedings of the 29th ACM International Conference on Multimedia*. p. 1600–1608. MM '21, Association for

- Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3474085.3475299>, <https://doi.org/10.1145/3474085.3475299>
26. Liu, R., Liu, Z., Liu, J., Fan, X., Luo, Z.: A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(10), 6594–6609 (2024). <https://doi.org/10.1109/TPAMI.2024.3382308>
 27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 9992–10002 (2021), <https://api.semanticscholar.org/CorpusID:232352874>
 28. Long, Y., Jia, H., Zhong, Y., Jiang, Y., Jia, Y.: Rxdnfnuse: A aggregated residual dense network for infrared and visible image fusion. *Information Fusion* **69**, 128–141 (2021). <https://doi.org/10.1016/j.inffus.2020.11.009>, <https://www.sciencedirect.com/science/article/pii/S1566253520304152>
 29. Ma, J., Liang, P., Yu, W., Chen, C., Guo, X., Wu, J., Jiang, J.: Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion* **54**, 85–98 (2020). <https://doi.org/10.1016/j.inffus.2019.07.005>, <https://www.sciencedirect.com/science/article/pii/S1566253519300314>
 30. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica* **9**(7), 1200–1217 (2022). <https://doi.org/10.1109/JAS.2022.105686>
 31. Ma, J., Tang, L., Xu, M., Zhang, H., Xiao, G.: Stdffusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Transactions on Instrumentation and Measurement* **70**, 1–13 (2021). <https://doi.org/10.1109/TIM.2021.3075747>
 32. Ma, J., Zhang, H., Shao, Z., Liang, P., Xu, H.: Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement* **70**, 1–14 (2021). <https://doi.org/10.1109/TIM.2020.3038013>
 33. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* **21**(12), 4695–4708 (2012). <https://doi.org/10.1109/TIP.2012.2214050>
 34. Petrovic, V., Xydeas, C.: Gradient-based multiresolution image fusion. *IEEE Transactions on Image Processing* **13**(2), 228–237 (2004). <https://doi.org/10.1109/TIP.2004.823821>
 35. Piella, G.: A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion* **4**(4), 259–280 (2003). [https://doi.org/10.1016/S1566-2535\(03\)00046-0](https://doi.org/10.1016/S1566-2535(03)00046-0), <https://www.sciencedirect.com/science/article/pii/S1566253503000460>
 36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: *Proceedings of the International Conference on Machine Learning* (2021)
 37. Rao, D., Xu, T., Wu, X.J.: Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing* pp. 1–1 (2023). <https://doi.org/10.1109/TIP.2023.3273451>
 38. Sun, Y., Cao, B., Zhu, P., Hu, Q.: Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology* pp. 1–1 (2022). <https://doi.org/10.1109/TCSVT.2022.3168279>

39. Tang, L., Deng, Y., Ma, Y., Huang, J., Ma, J.: Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica* **9**(12), 2121–2137 (2022). <https://doi.org/10.1109/JAS.2022.106082>
40. Tang, L., Li, C., Ma, J.: Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–18 (2025). <https://doi.org/10.1109/TPAMI.2025.3609323>
41. Tang, L., Li, C., Ma, J.: Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–18 (2025)
42. Tang, L., Yuan, J., Ma, J.: 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). <https://doi.org/10.1016/j.inffus.2021.12.004>, <https://www.sciencedirect.com/science/article/pii/S1566253521002542>
43. Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J.: Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion* **83–84**, 79–92 (2022). <https://doi.org/10.1016/j.inffus.2022.03.007>
44. Tang, W., He, F., Liu, Y.: Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia* **25**, 5413–5428 (2023). <https://doi.org/10.1109/TMM.2022.3192661>
45. Tang, W., He, F., Liu, Y., Duan, Y., Si, T.: Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology* **33**(7), 3159–3172 (2023). <https://doi.org/10.1109/TCSVT.2023.3234340>
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. p. 6000–6010 (2017)
47. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7464–7475 (2023). <https://doi.org/10.1109/CVPR52729.2023.00721>
48. Wang, Z., Zhang, Z., Qi, W., Yang, F., Xu, J.: Freqgan: Infrared and visible image fusion via unified frequency adversarial learning. *IEEE Transactions on Circuits and Systems for Video Technology* **35**(1), 728–740 (2025). <https://doi.org/10.1109/TCSVT.2024.3460172>
49. Wu, G., Liu, H., Fu, H., Peng, Y., Liu, J., Fan, X., Liu, R.: Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 17882–17891 (2025). <https://doi.org/10.1109/CVPR52734.2025.01666>
50. Xiao, Q., Jin, H., Su, H., Zhang, Y., Xiao, Z., Wang, B.: Spdfusion: A semantic prior knowledge-driven method for infrared and visible image fusion. *IEEE Transactions on Multimedia* **27**(1), 1691–1705 (2025). <https://doi.org/10.1109/TMM.2024.3521848>
51. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: *Neural Information Processing Systems (NeurIPS)* (2021)
52. Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H.: U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(1), 502–518 (2022). <https://doi.org/10.1109/TPAMI.2020.3012548>

53. Xu, H., Ma, J., Zhang, X.P.: Mef-gan: Multi-exposure image fusion via generative adversarial networks. *IEEE Transactions on Image Processing* **29**, 7203–7216 (2020). <https://doi.org/10.1109/TIP.2020.2999855>
54. Yang, B., Li, S.: Multifocus image fusion and restoration with sparse representation. *IEEE Transactions on Instrumentation and Measurement* **59**(4), 884–892 (2010). <https://doi.org/10.1109/TIM.2009.2026612>
55. Yang, K., Xiang, W., Chen, Z., Zhang, J., Liu, Y.: A review on infrared and visible image fusion algorithms based on neural networks. *Journal of Visual Communication and Image Representation* **101**, 104179 (2024). <https://doi.org/10.1016/j.jvcir.2024.104179>, <https://www.sciencedirect.com/science/article/pii/S1047320324001342>
56. Yi, X., Xu, H., Zhang, H., Tang, L., Ma, J.: Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 27016–27025 (2024). <https://doi.org/10.1109/CVPR52733.2024.02552>
57. Yi, X., Zhang, Y., Xiang, X., Yan, Q., Xu, H., Ma, J.: Lut-fuse: Towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2025)
58. Yuan, B., Sun, H., Guo, Y., Liu, Q., Zhan, X.: Explainable analysis of infrared and visible light image fusion based on deep learning. *Scientific Reports* **15**(1), 2223 (2025). <https://doi.org/10.1038/s41598-024-79684-6>, <https://doi.org/10.1038/s41598-024-79684-6>
59. Zhang, H., Zuo, X., Jiang, J., Guo, C., Ma, J.: Mrfs: Mutually reinforcing image fusion and segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 26964–26973 (2024). <https://doi.org/10.1109/CVPR52733.2024.02547>
60. Zhao, W., Cui, H., Wang, H., He, Y., Lu, H.: Freefusion: Infrared and visible image fusion via cross reconstruction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(9), 8040–8056 (2025). <https://doi.org/10.1109/TPAMI.2025.3572599>
61. Zhao, W., Xie, S., Zhao, F., He, Y., Lu, H.: Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13955–13965 (2023). <https://doi.org/10.1109/CVPR52729.2023.01341>
62. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5906–5916 (2023). <https://doi.org/10.1109/CVPR52729.2023.00572>
63. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Zhang, K., Xu, S., Chen, D., Timofte, R., Van Gool, L.: Equivariant multi-modality image fusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 25912–25921 (2024). <https://doi.org/10.1109/CVPR52733.2024.02448>
64. Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K., Meng, D., Timofte, R., Van Gool, L.: Ddfm: Denoising diffusion model for multi-modality image fusion. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8082–8093 (October 2023)
65. Zheng, N., Zhou, M., Huang, J., Hou, J., Li, H., Xu, Y., Zhao, F.: Probing synergistic high-order interaction in infrared and visible image fusion. In: *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26374–26385 (2024). <https://doi.org/10.1109/CVPR52733.2024.02492>