PCDFusion: Proposal-Context-Detail Bayesian Rendering for Infrared-Visible Image Fusion

Rui Liu[1] **Lin Gu**[2] **Hesong Li**[1] **Ying Fu**[1*]

[1]School of Computer Science and Technology,
Beijing Institute of Technology, Beijing 100081, China
[2]Research Institute of Electrical Communication,
Tohoku University, Sendai 980-8577, Japan

{liurui25,lihesong2,fuying}@bit.edu.cn
lin@tohoku.ac.jp

Abstract

Infrared-visible image fusion (IVIF) aims to combine visible and thermal infrared observations for human perception and downstream vision tasks. Visible and thermal infrared observations are physically complementary but unreliable in different ways. Existing deep IVIF methods mainly improve feature interaction and reconstruction, but often lack an explicit local rule for deciding when visible appearance should dominate and when thermal responses should contribute. To address this issue, we reformulate IVIF as local Bayesian posterior rendering, where fusion is controlled by posterior action weights rather than direct modality mixing. In this paper, we propose Proposal-Context-Detail Fusion (PCDFusion), a visible-anchored framework that implements this formulation through three rendering stages. PCDFusion preserves potentially useful thermal responses, updates exposure-degraded visible regions, and renders reliable band-limited infrared structures into the final luminance. Extensive experiments on public IVIF benchmarks show that PCDFusion improves fusion quality, target recovery, and downstream detection and segmentation performance under challenging illumination. The code is available at https://github.com/BIT-RuiLiu/PCDFusion.git.

1 Introduction

Infrared-visible image fusion (IVIF) aims to combine visible and thermal infrared observations into a unified image representation for human perception and downstream computer vision tasks. Visible and infrared images are physically heterogeneous sensing modalities. Visible cameras capture reflected radiance within the 0.4–$0.7\,\mu\mathrm{m}$ spectrum [1]. As illustrated in Fig. 1(a), visible images usually preserve natural appearance, rich textures, and fine structural details under proper exposure, but their quality strongly depends on lighting. In contrast, thermal infrared imaging mainly operates in the long-wave infrared band, approximately 8–$14\,\mu\mathrm{m}$, measuring emitted radiance related to object temperature [2, 3]. Thermal observations are therefore robust to darkness and illumination changes [4–7], but they often lack natural textures and reliable high-frequency structures.

Recent deep learning approaches improve IVIF through learnable representation [8–10], decomposition-based modeling [11, 12], attention-based interaction [13–15], generative reconstruction [16–18], and task-driven optimization [19–21]. However, most methods still formulate fusion as direct feature aggregation or modality balancing, implicitly assuming symmetric contributions from visible and infrared observations despite their physical asymmetry. As shown in Fig. 1(b), this

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

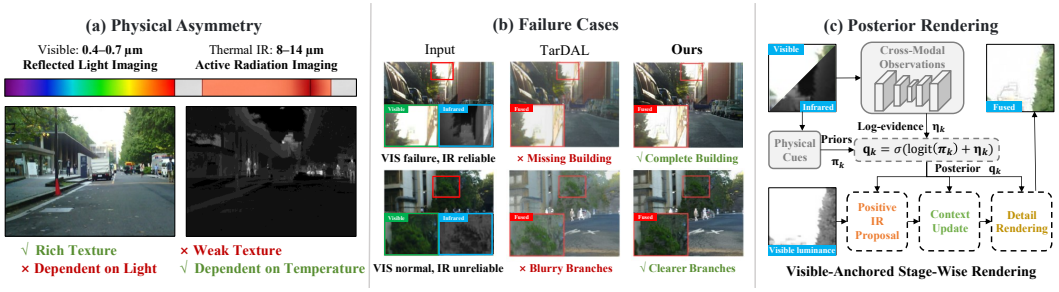


Figure 1: PCDFusion estimates posterior action weights from physical cues and cross-modal observations, recovering structures under visible failure while suppressing unreliable thermal responses.

limitation becomes evident under local reliability conflicts. When visible observations (green box) are saturated or degraded while infrared observations (blue box) retain useful structures, existing methods may still leave the failed visible region unresolved. For example, TarDAL [19] keeps the building region over-bright instead of recovering its structure from thermal observations. Conversely, when visible appearance remains reliable but infrared responses are unreliable for the local scene, direct fusion strategies tend to blur visible branches and introduce thermal contamination.

Motivated by this observation, we reformulate IVIF as local Bayesian posterior rendering under heterogeneous reliability. Instead of directly mixing visible and infrared features, each local rendering operation is treated as a posterior decision conditioned on physical cues and cross-modal observations. As illustrated in Fig. 1(c), physical cues define a reliability prior, while cross-modal observations provide a learned log-evidence score (log-odds correction) for the corresponding rendering action. The prior–evidence odds update then yields a normalized posterior action weight. Under this formulation, thermal responses are not uniformly injected into the visible-anchored rendering. They contribute where visible observations become unreliable and are suppressed where visible appearance remains reliable.

In this paper, we propose **P**roposal-**C**ontext-**D**etail Fusion (**PCDFusion**), a visible-anchored posterior rendering framework for IVIF. PCDFusion applies the same prior–evidence–posterior update rule across three sequential rendering stages. The proposal stage preserves potentially useful thermal responses as candidates for later rendering. The context stage updates visible regions degraded by exposure failure through posterior-weighted context recovery. The detail stage selectively renders locally reliable band-limited infrared structures into the final luminance. By separating thermal proposal, context correction, and detail rendering into different posterior decisions, PCDFusion reduces thermal contamination while preserving visible appearance and structural consistency. Extensive experiments on public IVIF benchmarks demonstrate improved fusion quality, target recovery, and downstream perception under challenging illumination. Our contributions are as follows:

- We formulate IVIF as local Bayesian posterior rendering, where physical priors and learned log-evidence scores determine reliability-conditioned rendering weights.
- We propose PCDFusion, a visible-anchored Proposal-Context-Detail network that keeps positive thermal responses available, updates exposure-degraded visible regions, and renders band-limited thermal structures into the final luminance.
- Without semantic labels for fusion training, PCDFusion improves image-level fusion quality and downstream detection and segmentation performance on public benchmarks.

2 Related Work

Learning-based IVIF approaches. Infrared-visible image fusion has progressed from hand-crafted fusion rules to learnable multimodal representation and reconstruction. Early deep models use autoencoders, convolutional networks, dense aggregation, residual fusion, or unified unsupervised mappings to combine visible and infrared inputs [8, 9, 22–26]. Decomposition-based methods further separate shared/specific or base/detail representations to improve structure preservation and modality complementarity [11, 12]. Recent methods enhance cross-modal interaction with Transformer-based architectures [13, 27, 28], cross-correlation learning [29], adaptive cross-attention [14, 30],

equivariant modeling [31], evolutionary search [32], degradation-robust denoising [33], diffusion-based generation [16, 17], and pretrained-knowledge-guided unified fusion [34]. Task-driven methods additionally optimize fused images for detection, segmentation, or semantic consistency, showing the importance of downstream perception [19, 35, 20, 36–39]. These methods improve feature extraction, visual reconstruction, and task utility, but they mainly focus on how visible and infrared information should be mixed. The local dominance relation between reflected-light appearance and thermal response remains mostly implicit.

Probabilistic and prior-guided fusion. Probabilistic modeling and explicit priors have also been explored in image fusion. Bayesian Fusion formulates IVIF as a hierarchical Bayesian regression problem and infers the fused image with EM optimization and total-variation regularization [40]. Diffusion-based fusion methods model multimodal fusion as conditional generation or denoising with task-specific fusion priors [16, 17]. Model-based Bayesian fusion networks further unfold hierarchical Bayesian inference into trainable neural architectures [41]. In parallel, prior-guided fusion methods introduce explicit observation cues to adapt fusion under adverse imaging conditions. Illumination-aware methods modulate fusion according to lighting states [42–44], while darkness-aware methods specifically compensate for degraded visible observations in low-light scenes [45]. These works validate probabilistic modeling and physical priors, but usually define uncertainty over the fused image or generation process, or use priors as auxiliary modulation. PCDFusion instead moves the probabilistic decision to the rendering level, estimating local posterior weights from physical priors and learned log-evidence scores.

3 PCDFusion

3.1 Formulation and Overview

Bayesian posterior formulation. Let $\mathbf{Y}_v$ and $\mathbf{Y}_r$ denote the visible luminance and thermal infrared observation, respectively. They are heterogeneous measurements: $\mathbf{Y}_v$ mainly reflects illumination-dependent appearance, while $\mathbf{Y}_r$ records emitted-radiance thermal responses. Therefore, fusion should not treat them as symmetric intensity sources. Instead, PCDFusion models each local rendering operation as a posterior decision conditioned on physical cues and cross-modal observations.

For the k-th rendering operation, we introduce a binary latent variable $z_k \in \{0, 1\}$, where $z_k = 1$ accepts the stage-specific rendering action and $z_k = 0$ retains its base outcome. Let $\mathcal{P}_k(\mathbf{Y}_v, \mathbf{Y}_r)$ denote the physical cue used to form the prior, and let $\mathcal{O}_k$ denote the local observations used for evidence estimation. An ideal probabilistic model defines

$$\boldsymbol{\pi}_k = p(z_k = 1 \mid \mathcal{P}_k), \qquad \boldsymbol{\ell}_k^+ = p(\mathcal{O}_k \mid z_k = 1, \mathcal{P}_k), \qquad \boldsymbol{\ell}_k^- = p(\mathcal{O}_k \mid z_k = 0, \mathcal{P}_k). \tag{1}$$

Bayes' rule in odds form gives

$$\frac{\mathbf{q}_k^\star}{1 - \mathbf{q}_k^\star} = \frac{\boldsymbol{\pi}_k}{1 - \boldsymbol{\pi}_k} \cdot \frac{\boldsymbol{\ell}_k^+}{\boldsymbol{\ell}_k^-}, \qquad \boldsymbol{\eta}_k^\star = \log \frac{\boldsymbol{\ell}_k^+}{\boldsymbol{\ell}_k^-}, \qquad \mathbf{q}_k^\star = \sigma(\mathrm{logit}(\boldsymbol{\pi}_k) + \boldsymbol{\eta}_k^\star)\,, \tag{2}$$

where $\mathbf{q}_k^\star = p(z_k = 1 \mid \mathcal{O}_k, \mathcal{P}_k)$ and $\boldsymbol{\eta}_k^\star$ is the exact log Bayes factor. In practice, PCDFusion does not fit the two class-conditional densities explicitly. Instead, it amortizes their relative evidence with a neural estimator:

$$\boldsymbol{\eta}_k = f_{\theta_k}(\mathcal{O}_k, \mathcal{P}_k), \qquad \mathbf{B}_k = \exp(\boldsymbol{\eta}_k), \qquad \mathbf{q}_k = \sigma(\mathrm{logit}(\boldsymbol{\pi}_k) + \boldsymbol{\eta}_k) = \frac{\boldsymbol{\pi}_k \mathbf{B}_k}{1 - \boldsymbol{\pi}_k + \boldsymbol{\pi}_k \mathbf{B}_k}. \tag{3}$$

Here, $\boldsymbol{\eta}_k$ is a learned log-evidence score (log-odds correction), and $\mathbf{B}_k$ is the corresponding multiplicative evidence factor. When $\boldsymbol{\eta}_k = 0$, the posterior remains at its prior value, $\mathbf{q}_k = \boldsymbol{\pi}_k$; positive and negative evidence shift the action odds above or below the physical prior. The resulting $\mathbf{q}_k$ is a normalized posterior action weight with an explicit prior anchor and a signed evidence correction.

The posterior is used to render the expectation of two stage-specific outcomes $\mathcal{R}_{k,0}$ and $\mathcal{R}_{k,1}$:

$$\widehat{\mathcal{R}}_k = \mathbb{E}[\mathcal{R}_k \mid \mathcal{O}_k, \mathcal{P}_k] = (1 - \mathbf{q}_k)\mathcal{R}_{k,0} + \mathbf{q}_k \mathcal{R}_{k,1}.$$

Proposal uses $\big(\mathbf{Y}_v,\ \mathbf{Y}_v + (1 - \mathbf{Y}_v) \odot \mathbf{G}_p\big)$, context uses $(\mathbf{Y}_p, \mathbf{Y}_{\mathrm{ctx}})$, and detail uses $(\mathbf{Y}_c, \mathbf{Y}_c + \mathbf{D}_r)$ as the two outcomes. The evidence estimators are optimized end-to-end through these posterior-weighted rendering actions, without direct probabilistic supervision of the scores. Calibration is

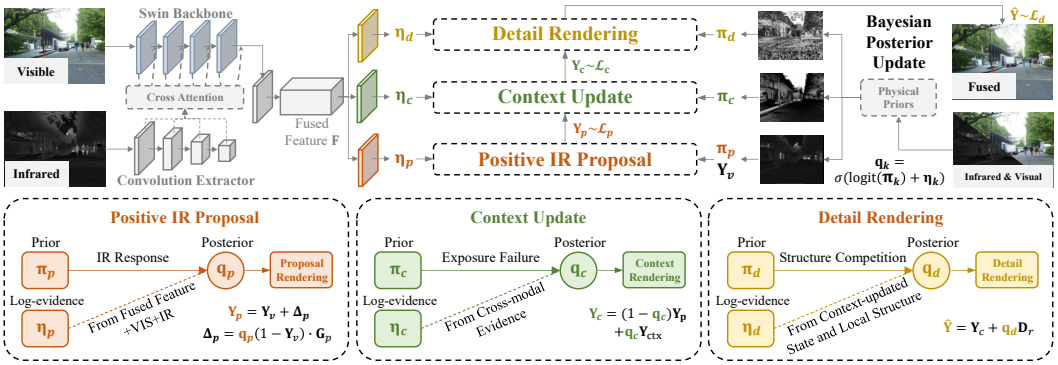


Figure 2: Overall architecture of the proposed PCDFusion.

evaluated empirically on the normalized posterior action weights. Together, the posterior update and outcome expectation form a local Bayesian rendering mechanism: $\boldsymbol{\pi}_k$ anchors the rendering decision, $\boldsymbol{\eta}_k$ adapts its log odds to local observations, and $\mathbf{q}_k$ mixes the base and action-conditioned outcomes. The same mechanism is instantiated at the proposal, context, and detail stages.

Architecture overview. As shown in Fig. 2, PCDFusion uses the visible image as the main appearance stream. A Swin Transformer extracts visible features, while a lightweight convolutional encoder extracts infrared features. The infrared features are injected into the visible stream through attention-based fusion, producing a visible-anchored representation $\mathbf{F}$. Based on Eq. (3), PCDFusion estimates three posterior action weights: $\mathbf{q}_p$ for positive thermal-response proposal, $\mathbf{q}_c$ for visible-context update, and $\mathbf{q}_d$ for infrared-detail rendering. The following subsections specify the corresponding priors, log-evidence estimators, and rendering equations.

3.2 Bayesian Positive Infrared Proposal

The first stage estimates whether positive thermal responses should be preserved before context and detail rendering. This stage is designed for high recall: potential thermal targets should remain available for later posterior decisions, but they should not be forced into the final rendering.

Prior and posterior. We use the normalized infrared observation to form a thermal-response prior:

$$\boldsymbol{\pi}_p = p(z_p = 1 \mid \mathcal{P}_p) = \sigma\left(\tau_p \left(\mathrm{Norm}(\mathbf{Y}_r) - 0.5\right)\right), \tag{4}$$

where $z_p = 1$ denotes that positive thermal responses should be preserved, and τ_p controls the prior sharpness. Since infrared intensity alone cannot determine whether the response should be retained, PCDFusion estimates the log-evidence score from the visible-anchored feature and the two observations:

$$\boldsymbol{\eta}_p = f_\theta^p(\mathbf{F}, \mathbf{Y}_v, \mathbf{Y}_r), \qquad \mathbf{q}_p = \sigma\big(\mathrm{logit}(\boldsymbol{\pi}_p) + \boldsymbol{\eta}_p\big). \tag{5}$$

Here, $\boldsymbol{\eta}_p$ is the learned log-evidence score and $\mathbf{q}_p$ is the proposal posterior action weight.

Proposal rendering. The proposal update should be positive and bounded by the remaining visible luminance margin. PCDFusion therefore predicts a conditional gain and renders the proposal as

$$\mathbf{G}_p = \sigma(g_\theta^p(\mathbf{F}, \mathbf{Y}_v, \mathbf{Y}_r)), \qquad \boldsymbol{\Delta}_p = \mathbf{q}_p \odot (1 - \mathbf{Y}_v) \odot \mathbf{G}_p, \qquad \mathbf{Y}_p = \mathbf{Y}_v + \boldsymbol{\Delta}_p. \tag{6}$$

Here, $\mathbf{G}_p$ is the conditional gain, $(1 - \mathbf{Y}_v)$ is the available positive luminance margin, and $\boldsymbol{\Delta}_p$ is the positive proposal correction.

3.3 Bayesian Visible Context Update

After obtaining $\mathbf{Y}_p$, PCDFusion estimates whether the visible-anchored context should be revised. This stage addresses illumination-induced visible failure: under-exposure lowers the effective signal-to-noise ratio, while over-exposure causes saturation and local contrast collapse.

Prior and posterior. The context prior is defined from visible luminance:

$$\boldsymbol{\pi}_c = p(z_c = 1 \mid \mathcal{P}_c) = 1 - 4\mathbf{Y}_v(1 - \mathbf{Y}_v), \tag{7}$$

where $z_c = 1$ indicates that the visible context should be updated. This U-shaped prior is low under normal exposure and high near darkness or saturation. Since exposure failure alone does not guarantee useful correction, PCDFusion estimates the log-evidence score from cross-modal observations and structural cues:

$$\boldsymbol{\eta}_c = f_\theta^c(\mathbf{F}, \mathbf{Y}_v, \mathbf{Y}_r, \boldsymbol{\pi}_p, \boldsymbol{\pi}_c, \mathbf{S}_v, \mathbf{S}_r^b), \quad \mathbf{q}_c = \sigma(\mathrm{logit}(\boldsymbol{\pi}_c) + \boldsymbol{\eta}_c)\,. \tag{8}$$

Here, $\mathbf{S}_v$ denotes visible structural support, $\mathbf{S}_r^b$ denotes band-limited infrared structural support, $\boldsymbol{\eta}_c$ is the learned log-evidence score, and $\mathbf{q}_c$ is the context posterior action weight.

Context rendering. The context action weight controls interpolation between the proposal rendering and a recovered context estimate. PCDFusion predicts the recovered context with a value prediction and a bounded residual refinement:

$$\mathbf{Y}_{\mathrm{ctx}} = \sigma(\mathbf{C}_\theta + \mathbf{R}_\theta)\,, \qquad \mathbf{Y}_c = (1 - \mathbf{q}_c) \odot \mathbf{Y}_p + \mathbf{q}_c \odot \mathbf{Y}_{\mathrm{ctx}}. \tag{9}$$

Here, $\mathbf{C}_\theta$ is the context value prediction, $\mathbf{R}_\theta$ is the bounded residual refinement, $\mathbf{Y}_{\mathrm{ctx}}$ is the recovered context, and $\mathbf{Y}_c$ is the context-updated rendering.

3.4 Bayesian Infrared Detail Rendering

After context update, PCDFusion estimates whether infrared detail should contribute to the final luminance. Raw infrared intensity is insufficient for this decision, because strong infrared responses may also come from smooth thermal background, emissivity variation, thermal reflection, or sensor artifacts. We therefore use local structure as the physical cue, so that infrared detail is favored only when it provides stronger local structural support than the visible anchor.

Prior and posterior. We compute visible structural support and infrared structural support as

$$\mathbf{S}_v = \mathrm{Norm}(|\nabla \mathbf{Y}_v|)\,, \qquad \mathbf{S}_r = \max(\mathrm{Norm}(|\nabla \mathbf{Y}_r|)\,, \mathrm{Norm}(\mathrm{Std}(|\mathcal{B}(\mathbf{Y}_r)|)))\,, \tag{10}$$

where $\mathcal{B}(\cdot)$ is a band-pass operator and $\mathrm{Std}(\cdot)$ computes local standard deviation. The detail prior compares the two structural supports:

$$\boldsymbol{\pi}_d = p(z_d = 1 \mid \mathcal{P}_d) = \frac{\mathbf{S}_r}{\mathbf{S}_r + \mathbf{S}_v + \epsilon}, \tag{11}$$

Since the structure prior alone cannot determine whether the current infrared detail is reliable for final rendering, PCDFusion further uses the context-updated state and local cross-modal observations to estimate the log-evidence score and posterior action weight:

$$\boldsymbol{\eta}_d = f_\theta^d(\mathbf{F}, \mathbf{Y}_v, \mathbf{Y}_r, \mathbf{Y}_c, \mathbf{S}_v, \mathbf{S}_r), \qquad \mathbf{q}_d = \sigma(\mathrm{logit}(\boldsymbol{\pi}_d) + \boldsymbol{\eta}_d)\,. \tag{12}$$

Detail rendering. To avoid transferring low-frequency thermal background into the final appearance, PCDFusion predicts a bounded residual and keeps only its band-limited component before posterior-weighted rendering:

$$\mathbf{D}_r = s_d \mathcal{B}\big(\tanh\big(d_\theta^d(\mathbf{F}, \mathbf{Y}_v, \mathbf{Y}_r, \mathbf{Y}_c)\big)\big)\,, \qquad \hat{\mathbf{Y}} = \mathbf{Y}_c + \mathbf{q}_d \odot \mathbf{D}_r. \tag{13}$$

Here, $\mathbf{D}_r$ is the band-limited infrared residual, s_d controls its scale, and $\hat{\mathbf{Y}}$ is the final fused luminance.

3.5 Loss Functions

PCDFusion is optimized with three losses for proposal recall, context posterior calibration, and final structural fidelity:

$$\mathcal{L} = \lambda_{\mathrm{p}} \mathcal{L}_{\mathrm{p}} + \lambda_{\mathrm{c}} \mathcal{L}_{\mathrm{c}} + \lambda_{\mathrm{d}} \mathcal{L}_{\mathrm{d}}, \tag{14}$$

where λ_{p}, λ_{c}, and λ_{d} weight the three loss terms, respectively.

Proposal response recall. The proposal rendering $\mathbf{Y}_p$ should retain possible thermal responses without forcing them into the final rendering. We derive an infrared-salient mask $\mathbf{M}_r$ from $\mathbf{Y}_r$ and define a relaxed thermal target and one-sided recall loss as

$$\mathbf{T}_r = \max\big(\mathbf{Y}_v - \boldsymbol{\pi}_c^2 \odot \mathrm{ReLU}(\mathbf{Y}_v - \mathbf{Y}_r), \mathbf{Y}_r\big)\,, \qquad \mathcal{L}_p = \frac{\|\mathbf{M}_r \odot \mathrm{ReLU}(\mathbf{T}_r - \mathbf{Y}_p)\|_1}{\|\mathbf{M}_r\|_1 + \epsilon}. \tag{15}$$

Context posterior calibration. The context action weight should activate only where unreliable visible observation and infrared structure jointly support context update. We use a weak calibration target and a keep term for reliable visible regions:

$$\widetilde{\mathbf{q}}_c = \mathrm{sg}\left(\boldsymbol{\pi}_c \odot \mathbf{S}_r\right), \qquad \mathcal{L}_c = \left\|\mathbf{q}_c - \widetilde{\mathbf{q}}_c\right\|_1 + \lambda_{\mathrm{keep}} \left\|\mathrm{sg}(1-\boldsymbol{\pi}_c) \odot \mathbf{q}_c\right\|_1, \tag{16}$$

where $\mathrm{sg}(\cdot)$ stops gradients through the calibration target and the reliable-region mask.

Final structural fidelity. The final rendering should preserve the strongest local structure without direct intensity forcing. We therefore match the gradient magnitude of $\hat{\mathbf{Y}}$ to the element-wise maximum gradient response:

$$\mathcal{L}_d = \left\| |\nabla \hat{\mathbf{Y}}| - \max(|\nabla \mathbf{Y}_v|, |\nabla \mathbf{Y}_r|) \right\|_1. \tag{17}$$

4 Experiments

4.1 Settings and Details

We evaluate PCDFusion on three public infrared-visible benchmarks: MSRS [42], FMB [20], and M3FD [19]. The fusion model is trained only on the MSRS training split and directly evaluated on all datasets without dataset-specific fine-tuning. Following common IVIF evaluation protocols [4, 5], we report AG, SCD [46], VIF [47], Qabf [48], and Qcb [49]. For downstream perception, we use SegFormerB2 [50] for semantic segmentation and YOLOv10s [51] for object detection. For each fusion method, the corresponding fused training images are used to fine-tune the task model from the same pretrained initialization under identical settings, and the fine-tuned model is then evaluated on the fused test images. We report mIoU on FMB and $\mathrm{mAP}_{50:95}$ on M3FD. PCDFusion is implemented in PyTorch and trained for 200 epochs using AdamW with an initial learning rate of 2×10^{-4}, weight decay 10^{-4}, batch size 16, crop size 256×256, and cosine learning-rate decay. During inference, the final RGB image is reconstructed by combining the fused luminance with the visible chrominance.

4.2 Results of Infrared-Visible Image Fusion

We compare PCDFusion with recent representative IVIF methods, including SwinFusion [13], TarDAL [19], CDDFuse [12], Diff-IF [17], EMMA [31], DCEvo [32], Deno-IF [33], SAGE [37], and UPFusion [34].

Quantitative comparisons. Tab. 1 reports quantitative fusion results on MSRS, FMB, and M3FD. PCDFusion shows consistent advantages in preserving visual information, transferring edge structures, and maintaining perceptual quality, as reflected by its strong VIF, Qabf, and Qcb results. On MSRS, it combines the highest image sharpness with leading information-fidelity and edge-preservation scores, suggesting that enhanced local details are accompanied by effective retention of source content. This advantage extends to FMB and M3FD, where PCDFusion achieves the best edge-based and perceptual-quality scores while maintaining competitive sharpness. For example, its Qabf on M3FD reaches 0.67, compared with 0.64 for the strongest baseline, supporting more effective transfer of structural boundaries into the fused image. Although some competing methods achieve higher AG or SCD, these gains are not accompanied by comparable improvements in information fidelity and perceptual quality. Taken together, the results suggest that PCDFusion balances detail enhancement with structural preservation.

Qualitative comparisons. Fig. 3 shows representative MSRS and FMB scenes with local zoom-ins. In the first scene, PCDFusion restores clearer building structures in the saturated visible region without contaminating the surrounding sky. In the second scene, the visible image remains reliable, while the infrared observation fails around the glass region. Competing methods tend to inject unreliable thermal responses or disturb the visible appearance, whereas PCDFusion suppresses the invalid infrared response and preserves the original structure. In the third scene, haze and low contrast make thin structures difficult to preserve. PCDFusion maintains clearer utility-pole structures while avoiding excessive thermal amplification in the background. These results show that PCDFusion can both recover useful thermal structures under visible failure and reject unreliable infrared responses when visible appearance is reliable.

Table 1: Quantitative fusion results on the MSRS, FMB, and M3FD datasets. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	MSRS					FMB					M3FD				
	AG↑	SCD↑	VIF↑	Qabf↑	Qcb↑	AG↑	SCD↑	VIF↑	Qabf↑	Qcb↑	AG↑	SCD↑	VIF↑	Qabf↑	Qcb↑
SwinFusion JAS'22	3.56	1.69	1.03	0.65	0.58	4.05	1.56	1.03	0.66	0.46	4.61	1.56	0.90	0.61	0.46
TarDAL CVPR'22	2.93	1.37	0.72	0.42	0.27	3.54	1.57	0.81	0.46	0.40	4.26	1.52	0.71	0.43	0.41
CDDFuse CVPR'23	3.75	1.62	1.08	0.68	0.57	4.31	1.62	1.06	0.67	0.47	4.86	1.65	0.92	0.61	0.46
Diff-IF Inf. Fusion'24	3.70	1.62	1.10	0.69	0.59	4.05	1.37	1.05	0.66	0.46	4.57	1.32	0.89	0.59	0.45
EMMA CVPR'24	3.79	1.63	1.02	0.63	0.56	**4.68**	1.50	1.03	0.64	0.46	**5.33**	1.49	0.90	0.59	0.46
DCEvo CVPR'25	3.72	1.66	1.07	0.69	0.59	4.05	1.50	1.06	0.69	0.47	4.62	1.49	0.92	0.63	0.47
Deno-IF NIPS'25	3.11	**1.79**	0.90	0.56	0.54	3.19	1.65	0.83	0.57	0.42	3.71	1.68	0.74	0.53	0.42
SAGE CVPR'25	3.16	1.42	0.76	0.54	0.50	3.78	**1.71**	0.97	0.64	0.45	4.25	**1.69**	0.86	0.58	0.45
UPFusion AAAI'26	3.56	1.61	1.01	0.64	0.58	4.10	1.43	**1.08**	0.68	0.47	4.70	1.39	0.95	0.64	0.47
PCDFusion Ours	**3.84**	1.67	**1.12**	**0.72**	**0.59**	4.33	1.57	1.08	**0.71**	**0.49**	5.04	1.58	**0.95**	**0.67**	**0.50**



Figure 3: Visual comparisons of infrared-visible fusion results on MSRS and FMB dataset scenes.

4.3 Results of Downstream Tasks

Tab. 2 reports downstream results on FMB segmentation and M3FD detection. PCDFusion achieves the best aggregate mIoU and $mAP_{50:95}$, although no method ranks first for every class. Its 60.05% mIoU exceeds Diff-IF (58.98%) by 1.07 percentage points, and its 55.98% $mAP_{50:95}$ exceeds DCEvo (55.79%) by 0.19 percentage points.

Fig. 4 shows representative downstream visualizations. For segmentation, PCDFusion produces a more complete pole shape, while competing methods often fragment or miss thin structures. For detection, PCDFusion avoids falsely detecting the bright region on the far right and maintains reliable localization of valid targets. These results indicate cleaner structure preservation and less background-induced task interference.

Table 2: Downstream perception results on FMB semantic segmentation and M3FD object detection.

Method	FMB Segmentation							M3FD Detection						
	Side.	Light	Sign	Car	Truck	Bus	mIoU ↑	Per.	Car	Bus	Motor.	Lamp	Truck	$mAP_{50:95}$ ↑
SwinFusion JAS'22	48.98	43.22	80.68	84.23	48.77	57.57	58.33	48.90	68.13	71.31	40.82	42.03	58.00	54.86
TarDAL CVPR'22	50.05	42.30	81.20	81.85	35.82	26.36	55.59	48.80	66.51	69.55	34.00	38.87	56.38	52.35
CDDFuse CVPR'23	50.97	40.85	81.80	82.90	22.14	55.70	56.66	48.79	68.44	71.80	38.00	43.44	54.39	54.15
Diff-IF Inf. Fusion'24	43.81	45.24	82.04	83.57	45.22	**75.99**	58.98	49.77	67.43	70.08	38.47	39.82	58.68	54.04
EMMA CVPR'24	49.32	**49.42**	81.24	82.58	42.99	59.41	58.47	48.35	68.03	71.02	38.26	43.70	57.20	54.43
DCEvo CVPR'25	48.01	46.98	81.01	**84.55**	43.67	49.08	57.73	**50.05**	68.25	71.60	41.20	43.86	**59.78**	55.79
Deno-IF NIPS'25	47.39	40.84	80.57	82.25	40.13	34.40	55.45	50.03	68.68	70.37	41.06	43.16	57.61	55.15
SAGE CVPR'25	**51.65**	42.90	**82.37**	82.28	41.08	31.51	56.33	49.59	68.42	68.91	39.15	43.22	56.39	54.28
UPFusion AAAI'26	43.86	19.68	77.97	81.92	39.27	58.29	55.36	48.72	68.52	71.28	38.57	**45.94**	58.04	55.18
PCDFusion Ours	50.79	45.89	81.31	84.44	**51.54**	71.19	**60.05**	48.95	**69.00**	**72.60**	**42.89**	44.53	57.93	**55.98**

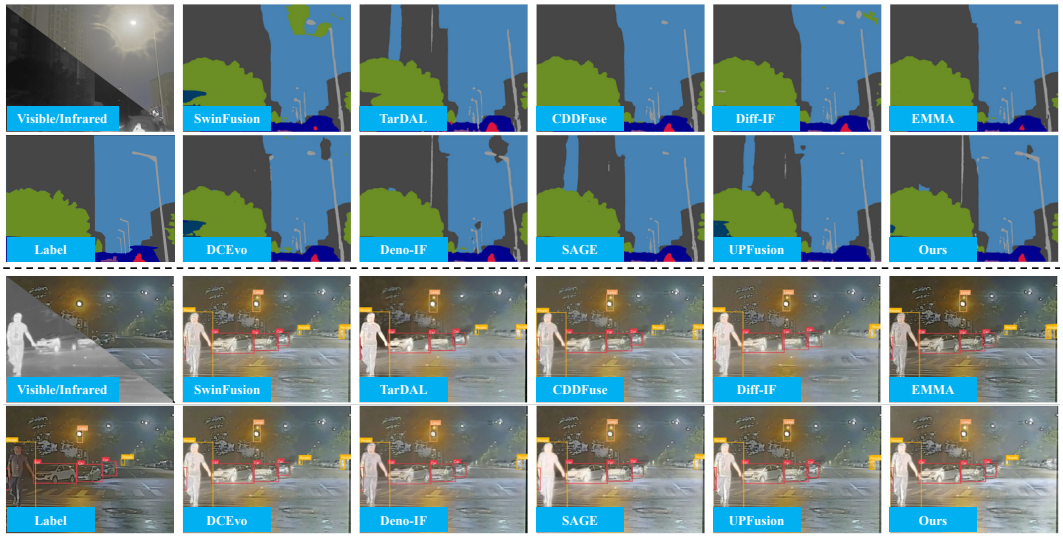


Figure 4: Visualization results on downstream perception tasks.

Table 3: Ablation of PCDFusion stages.

Method	AG↑	SCD↑	VIF↑	Qabf↑	Qcb↑	mIoU↑	mAP↑
w/o P	3.810	1.594	1.116	0.723	0.592	58.44	55.18
w/o C	3.825	**1.716**	1.107	0.719	0.582	58.86	54.88
w/o D	**3.846**	1.618	1.110	**0.723**	0.592	58.89	**56.13**
Full	3.839	1.671	**1.117**	0.723	**0.593**	**60.05**	55.98

Table 4: Ablation of posterior estimation forms.

Method	AG↑	SCD↑	VIF↑	Qabf↑	Qcb↑	mIoU↑	mAP↑
only Prior	3.789	1.666	**1.123**	**0.728**	**0.597**	59.09	55.56
only Evidence	**3.885**	**1.696**	1.104	0.722	0.589	59.04	55.38
Product	3.835	1.596	1.102	0.722	0.588	58.89	55.76
Full	3.839	1.671	1.117	0.723	0.593	**60.05**	**55.98**

4.4 Discussions

To analyze the effectiveness of PCDFusion, we conduct ablation studies on MSRS.

Effectiveness of different rendering stages. Tab. 3 and the stage outputs in the left of Fig. 5 show that the three stages perform complementary rendering functions. Removing P causes the most evident degradation, confirming that early preservation of positive thermal responses is necessary. Removing C weakens reliable reconstruction despite increasing SCD, indicating that context recovery cannot be driven only by cross-modal variation. Removing D reduces the final balance, showing that band-limited thermal detail still requires selective rendering. $\mathbf{Y}_p$ incorporates infrared targets such as pedestrians and vehicles, but still misses details in the over-exposed building region; $\mathbf{Y}_c$ restores the building structure in the visible-failure area; and $\hat{\mathbf{Y}}$ further sharpens the recovered details.

Effectiveness of Bayesian priors and posteriors. Tabs. 4 and 5 analyze the Bayesian design. Using only priors yields strong image-level fidelity, confirming that the physical cues are informative; using only learned evidence improves AG and SCD but weakens downstream performance, suggesting over-amplified local variations. The product form is less stable than the normalized posterior, while



Figure 5: Left: luminance and RGB outputs of the proposal, context, and detail rendering stages. Right: priors, learned log-evidence scores, and posterior action weights of the three Bayesian processes.

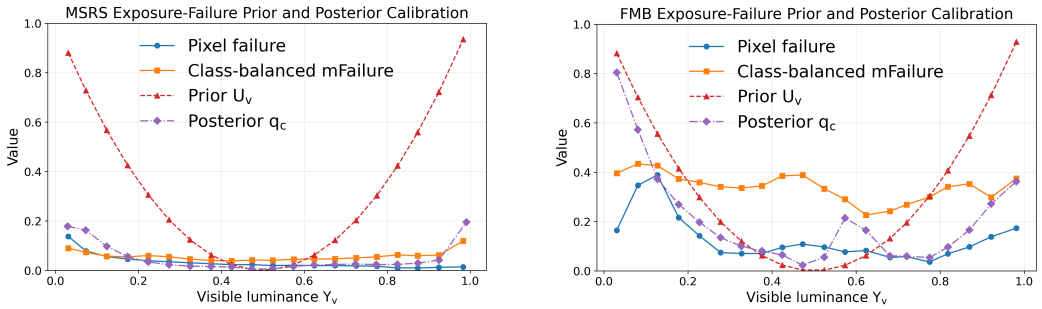


Figure 6: Calibration of the exposure-failure prior and posterior context update weight.

the full update gives the best downstream results with balanced fusion quality. For context modeling, one-sided or linear exposure priors are less effective than the adopted symmetric exposure-failure prior, especially on VIF, Qcb, mIoU, and mAP. The Bayesian maps in the right of Fig. 5 show that the posteriors refine their priors: $\mathbf{q}_p$ strengthens reliable thermal-response regions, $\mathbf{q}_c$ restricts context update to unreliable visible areas, and $\mathbf{q}_d$ localizes final detail compensation. In Fig. 6, $\mathbf{q}_c$ also follows the empirical exposure-failure trends on MSRS and FMB, indicating good posterior calibration across datasets.

Effectiveness of loss functions. Tab. 6 validates the objective design. Removing $\mathcal{L}_p$ weakens proposal recall and reduces downstream performance, showing that early thermal-response preservation requires explicit supervision. Removing $\mathcal{L}_c$ or the keep term can improve some image-level scores, but lowers mIoU or mAP, indicating that context posterior calibration is necessary for stable task utility. Removing $\mathcal{L}_d$ causes a severe collapse in Qabf, Qcb, mIoU, and mAP, confirming that final structural fidelity is indispensable. Directly applying saliency supervision to the final output increases SCD, but weakens VIF, Qabf, Qcb, and downstream results, supporting our choice to supervise thermal responses at the proposal stage rather than force them into the final rendering.

Robustness under input corruption. Tab. 7 evaluates PCDFusion under both single- and dual-modal corruption. Across the 15 corruption settings, PCDFusion ranks first in 11 and second in the remaining four, while consistently outperforming UPFusion. When only one modality is degraded, PCDFusion achieves the highest mIoU in eight of nine settings, indicating that the posterior rendering mechanism can exploit reliable complementary information from the other modality. Under joint corruption, it remains the best method for both severity-1 settings and the severity-2 VIS-dark setting, and maintains higher mIoU than UPFusion in all six cases. Although SAGE performs better under several severe joint corruptions, PCDFusion exhibits a more gradual performance decrease as the corruption strength increases. These results demonstrate that the physical priors and observation-dependent posterior weights effectively adapt the contribution of each modality to its local reliability, improving robustness over fixed fusion strategies.

Table 5: Ablation of context prior estimates.

Priors	AG↑	SCD↑	VIF↑	Qabf↑	Qcb↑	mIoU↑	mAP↑
$1-\mathbf{Y}_v$	3.839	**1.693**	1.111	**0.723**	0.589	58.88	55.53
$\mathbf{Y}_v$	**3.875**	1.692	1.106	0.722	0.588	59.19	55.46
$\lvert 2\mathbf{Y}_v-1\rvert$	3.863	1.675	1.107	0.721	0.587	58.94	55.50
$(2\mathbf{Y}_v-1)^2$	3.839	1.671	**1.117**	**0.723**	**0.593**	**60.05**	**55.98**

Table 6: Ablation of objectives.

Method	AG↑	SCD↑	VIF↑	Qabf↑	Qcb↑	mIoU↑	mAP↑
w/o $\mathcal{L}_p$	3.868	1.658	1.105	0.721	0.589	59.02	56.00
w/o $\mathcal{L}_c$	**3.872**	1.678	1.111	0.723	**0.594**	58.90	55.58
w/o keep	3.862	1.675	1.107	0.722	0.586	58.79	**56.01**
w/o $\mathcal{L}_d$	3.798	1.070	1.010	0.178	0.332	41.01	23.83
Final Sal.	3.852	**1.735**	1.098	0.720	0.582	58.56	55.15
Full	3.839	1.671	**1.117**	**0.723**	0.593	**60.05**	55.98

Table 7: FMB mIoU under input corruption. Slash-separated values follow severities 1/2/3.

Corruption	Severity	UPFusion	SAGE	PCDFusion
Visible dark	1/2/3	55.93 / 51.75 / 47.37	56.72 / 53.79 / 51.02	**59.75 / 57.09 / 52.14**
Visible bright	1/2/3	54.97 / 50.61 / 45.98	56.86 / 55.33 / **51.28**	**59.84 / 56.07** / 50.28
Infrared only	1/2/3	54.62 / 53.76 / 53.10	55.94 / 55.43 / 54.55	**58.06 / 57.44 / 56.54**
Both (VIS dark)	1/2/3	54.05 / 50.41 / 42.90	56.00 / 53.38 / **50.06**	**57.66 / 53.45** / 44.23
Both (VIS bright)	1/2/3	53.92 / 48.75 / 43.32	56.23 / **54.27 / 47.85**	**58.46** / 53.94 / 45.83

Table 8: Ablation of rendering order and joint decisions.

Order	AG↑	VIF↑	Qabf↑	mIoU↑	mAP↑
P–C–D	**3.839**	**1.117**	**0.723**	**60.05**	**55.98**
C–P–D	3.820	1.091	0.715	56.30	54.86
P–D–C	3.792	1.110	0.720	56.24	55.68
Parallel	3.824	1.107	0.718	56.49	55.33

Sequential versus joint rendering decisions. Tab. 8 evaluates whether the three rendering decisions should be performed sequentially or jointly. The proposed P–C–D sequence achieves the best result on every fusion and downstream metric. Compared with the strongest alternative for each metric, it improves AG, VIF, Qabf, mIoU, and mAP by 0.015, 0.007, 0.003, 3.56, and 0.30, respectively. In particular, placing context update before proposal rendering reduces mIoU from 60.05 to 56.30, while applying detail rendering before context update reduces it to 56.24. The parallel joint-gating variant also reaches only 56.49 mIoU, showing that the improvement cannot be explained merely by introducing three rendering decisions. Instead, the stages benefit from their ordered dependency: proposal rendering first preserves candidate thermal responses, context rendering then corrects the visible representation using this proposal, and detail rendering finally injects structural residuals into the corrected representation. The consistent gains across image-level quality and downstream performance validate the proposed sequential posterior-rendering design.

Limitations. PCDFusion still has limitations in low-luminance and infrared-unreliable regions. As shown in Fig. 1(b), although it reduces the influence of unreliable infrared responses, the tree branches remain slightly blurred. In the second scene of Fig. 3, the glass region is largely invalid in infrared, but minor thermal interference is still transferred into the fused result. A possible reason is that the context posterior in Fig. 6 remains relatively high in low-luminance ranges, which helps recover degraded visible regions but can also make the model over-permissive when infrared responses are unreliable. Future work will improve posterior calibration with stronger structure-preservation constraints and more explicit infrared reliability estimation.

5 Conclusion

In this paper, we presented PCDFusion, a visible-anchored posterior rendering framework for IVIF. PCDFusion formulates fusion as local reliability-conditioned rendering from heterogeneous visible and thermal observations. It uses proposal, context, and detail rendering to preserve positive thermal responses, update exposure-degraded visible regions, and inject band-limited infrared structures. By combining physical priors with learned log-evidence scores, PCDFusion selectively uses thermal responses while preserving reliable visible appearance. Experiments on public benchmarks and downstream tasks show improved fusion quality, target preservation, and task utility. Nevertheless, low-luminance regions and infrared-unreliable cases remain challenging, especially when weak visible structures should be preserved while invalid thermal responses should be suppressed. Future work will explore stronger posterior calibration and more explicit infrared reliability estimation.

6 Acknowledgments

This work was supported by the National NaturalScience Foundation of China (62671049 and 62331006), and Beijing Natural Science Foundation (L253009).

References

[1] Schowengerdt, R. A. *Remote Sensing: Models and Methods for Image Processing*. Academic Press, Burlington, MA, 3rd edn., 2007.

[2] Vollmer, M., K.-P. Möllmann. *Infrared Thermal Imaging: Fundamentals, Research and Applications*. Wiley-VCH, Weinheim, Germany, 2nd edn., 2018.

[3] Kuenzer, C., S. Dech, eds. *Thermal Infrared Remote Sensing: Sensors, Methods, Applications*, vol. 17 of *Remote Sensing and Digital Image Processing*. Springer Netherlands, Dordrecht, 2013.

[4] Ma, J., Y. Ma, C. Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019.

[5] Zhang, X., Y. Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023.

[6] Ma, W., K. Wang, J. Li, et al. Infrared and visible image fusion technology and application: A review. *Sensors*, 23(2):599, 2023.

[7] Zhang, Y., Z. Lai, T. Zhang, et al. Unaligned RGB guided hyperspectral image super-resolution with spatial-spectral concordance. *International Journal of Computer Vision*, 133(9):6590–6610, 2025.

[8] Li, H., X.-J. Wu. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

[9] Li, H., X.-J. Wu, J. Kittler. RFN-Nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021.

[10] Liu, Q., R. Liu, B. Zheng, et al. Infrared small target detection with scale and location sensitivity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17490–17499. IEEE, 2024.

[11] Zhao, Z., S. Xu, C. Zhang, et al. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 970–976. 2020.

[12] Zhao, Z., H. Bai, J. Zhang, et al. 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*, pages 5906–5916. 2023.

[13] Ma, J., L. Tang, F. Fan, et al. SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022.

[14] Park, I., Y. Jin, Y. Nalcakan, et al. IVIFormer: Illumination-aware infrared-visible image fusion via adaptive domain-switching cross attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 2241–2250. 2025.

[15] Liu, R., J. Nie, Y. Fu. RDANet: Relative degradation aware network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

[16] Zhao, Z., H. Bai, Y. Zhu, et al. DDFM: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093. 2023.

[17] Yi, X., L. Tang, H. Zhang, et al. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.

[18] Zhang, Y., S. Chen, Y. Tian, et al. Supervise-assisted multi-modality fusion diffusion model for PET restoration. *Tsinghua Science and Technology*, 2026.

[19] Liu, J., X. Fan, Z. Huang, et al. 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*, pages 5802–5811. 2022.

[20] Liu, J., Z. Liu, G. Wu, et al. 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*, pages 8115–8124. 2023.

[21] Zhu, T., H. Li, Y. Fu. TriM-SOD: A multi-modal, multi-task, and multi-scale spacecraft optical dataset. *Space: Science & Technology*, 5:0299, 2025.

[22] Xu, H., J. Ma, J. Jiang, et al. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022.

[23] Zhang, T., B. Wang, Y. Fu, et al. Deep learning-based spectral image super-resolution: A survey. *Journal of Image and Graphics*, 29(8):2113–2136, 2024.

[24] Li, H., Y. Fu, Z. Wu, et al. Spatial-frequency learning for scanning transmission electron microscopy image enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2026.

[25] Zhu, T., Y. Fu, H. Li, et al. Binarized high-efficiency RAW video restoration and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2026.

[26] Li, H., Y. Fu. FCDFusion: A fast, low color deviation method for fusing visible and infrared image pairs. *Computational Visual Media*, 11(1):195–211, 2025.

[27] Li, M., Y. Fu. Joint self-attention Transformer for multispectral and hyperspectral image fusion. *Journal of Image and Graphics*, 28(12):3922–3934, 2023.

[28] Tian, Y., Y. Fu, J. Zhang. Transformer-based under-sampled single-pixel imaging. *Chinese Journal of Electronics*, 32(5):1151–1159, 2023.

[29] Tang, W., F. He, Y. Liu. TCCFusion: An infrared and visible image fusion method based on Transformer and cross correlation. *Pattern Recognition*, 137:109295, 2023.

[30] Zhang, T., Y. Fu, J. Zhang. Deep guided attention network for joint denoising and demosaicing in real image. *Chinese Journal of Electronics*, 33(1):303–312, 2024.

[31] Zhao, Z., H. Bai, J. Zhang, et al. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25912–25921. 2024.

[32] Liu, J., B. Zhang, Q. Mei, et al. 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*, pages 2226–2235. 2025.

[33] Xu, H., Y. Li, Y. Deng, et al. Deno-IF: Unsupervised noisy visible and infrared image fusion method. In *Advances in Neural Information Processing Systems*, vol. 38, pages 141649–141671. 2025.

[34] Li, X., X. Li, W. Jiang. Text-guided channel perturbation and pre-trained knowledge integration for unified multi-modality image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, pages 6521–6529. 2026.

[35] Liu, R., Z. Liu, J. Liu, et al. 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.

[36] Zhang, H., X. Zuo, J. Jiang, et al. MRFS: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26974–26983. 2024.

[37] Wu, G., H. Liu, H. Fu, et al. 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*, pages 17882–17891. 2025.

[38] Liu, J., G. Wu, Z. Liu, et al. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025.

[39] Chen, L., Y. Fu, J. Shangguan, et al. A survey on degraded image segmentation. *Chinese Journal of Electronics*, 35(5):1–21, 2026.

[40] Zhao, Z., S. Xu, C. Zhang, et al. Bayesian fusion for infrared and visible images. *Signal Processing*, 177:107734, 2020.

[41] Li, W., Y. Kuang, Z. Cai, et al. Model-based Bayesian Fusion-Net for infrared and visible image fusion. *EURASIP Journal on Image and Video Processing*, 2025(1):14, 2025.

[42] Tang, L., J. Yuan, H. Zhang, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022.

[43] Yang, Q., Y. Zhang, Z. Zhao, et al. IAIFNet: An illumination-aware infrared and visible image fusion network. *IEEE Signal Processing Letters*, 31:1374–1378, 2024.

[44] Ding, S., J. Huang, Z. Cai, et al. FIAFusion: A feedback-based illumination-adaptive infrared and visible image fusion method. *IEEE Sensors Journal*, 25(4):7667–7680, 2025.

[45] Tang, L., X. Xiang, H. Zhang, et al. DIVFusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023.

[46] Aslantas, V., E. Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU - International Journal of Electronics and Communications*, 69(12):1890–1896, 2015.

[47] Sheikh, H. R., A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

[48] Xydeas, C. S., V. Petrovic. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000.

[49] Chen, Y., R. S. Blum. A new automated quality assessment algorithm for image fusion. *Image and Vision Computing*, 27(10):1421–1432, 2009.

[50] Xie, E., W. Wang, Z. Yu, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, vol. 34, pages 12077–12090. 2021.

[51] Wang, A., H. Chen, L. Liu, et al. YOLOv10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems*, vol. 37, pages 107984–108011. 2024.

Supplementary Material

A Implementation Details

Network Implementation. PCDFusion follows the visible-anchored feature flow described in the main paper. The visible branch uses a single-scale Swin Transformer stream with patch size 8, embedding dimension 192, and depth 4. The four Swin blocks are divided into two stages. The token resolution is kept at $H/8 \times W/8$ throughout the visible backbone.

The infrared branch is implemented as a lightweight convolutional encoder with base width 24. It extracts multi-scale infrared features at resolutions $H/2$, $H/4$, $H/8$, and $H/16$, with channel widths 48, 96, 192, and 192, respectively. All infrared features are projected to the visible token dimension and concatenated as infrared tokens $\mathbf{T}_r$. Let $\mathbf{T}_v$ denote the visible tokens. The infrared-to-visible injection is implemented by cross-attention:

$$\mathbf{Q} = \mathbf{T}_v \mathbf{W}_Q, \qquad \mathbf{K} = \mathbf{T}_r \mathbf{W}_K, \qquad \mathbf{V} = \mathbf{T}_r \mathbf{W}_V, \tag{18}$$

$$\widetilde{\mathbf{T}}_v = \mathbf{T}_v + \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{19}$$

Here, visible tokens provide queries, infrared tokens provide keys and values, and d is the attention dimension.

After cross-modal injection, the visible token stream is unembedded back to the image resolution. The recovered visible-anchored feature is concatenated with the visible image and the infrared observation, and then mapped to a full-resolution hidden feature with width 48. The proposal, context, and detail rendering heads are applied on this hidden feature to predict the posterior weights and rendering terms used in the main paper.

Table 9: Probabilistic roles and their realizations in PCDFusion.

Role	Probabilistic operation	PCDFusion realization	Consequence
Decision	Bernoulli rendering action z_k	Explicit proposal/context/detail action semantics	Defined local inference target
Prior	$p(z_k = 1 \mid \mathcal{P}_k)$	Analytic IR-response, exposure, or structure cue	Physical zero-evidence reference
Evidence	Relative support in log Bayes-factor coordinates	Neural score $f_{\theta_k}(\mathcal{O}_k, \mathcal{P}_k)$	Amortized log-evidence correction
Update	Prior odds $\times$ evidence factor	$\mathbf{q}_k = \sigma(\mathrm{logit}(\boldsymbol{\pi}_k) + \boldsymbol{\eta}_k)$	Normalized action posterior
Rendering	Expectation over two actions	$(1 - \mathbf{q}_k)\mathcal{R}_{k,0} + \mathbf{q}_k \mathcal{R}_{k,1}$	Posterior-weighted fusion
Learning	Shared estimator of observation evidence	End-to-end optimization of θ_k through rendering losses	Amortized across pixels and images

Tab. 9 summarizes how the probabilistic quantities map to the implemented rendering pipeline. The binary variable specifies the local action being inferred, the analytic cue provides its physical prior, and the neural estimator supplies an observation-dependent correction in log-odds coordinates. The normalized posterior then mixes the base and action-conditioned outcomes, so all three rendering stages share the same inference structure while using different physical cues and outcomes.

RGB Reconstruction. PCDFusion predicts the fused luminance $\hat{\mathbf{Y}}$. Given the visible RGB image $\mathbf{I}_{\mathrm{v}} = (\mathbf{R}_{\mathrm{v}}, \mathbf{G}_{\mathrm{v}}, \mathbf{B}_{\mathrm{v}})$, we first compute its luminance and chrominance channels by

$$\mathbf{Y}_{\mathrm{v}} = 0.299\mathbf{R}_{\mathrm{v}} + 0.587\mathbf{G}_{\mathrm{v}} + 0.114\mathbf{B}_{\mathrm{v}}, \tag{20}$$

$$\mathbf{Cb}_{\mathrm{v}} = -0.168736\mathbf{R}_{\mathrm{v}} - 0.331264\mathbf{G}_{\mathrm{v}} + 0.5\mathbf{B}_{\mathrm{v}} + 0.5, \tag{21}$$

$$\mathbf{Cr}_{\mathrm{v}} = 0.5\mathbf{R}_{\mathrm{v}} - 0.418688\mathbf{G}_{\mathrm{v}} - 0.081312\mathbf{B}_{\mathrm{v}} + 0.5. \tag{22}$$

The final RGB image is reconstructed by replacing the visible luminance with the predicted fused luminance $\hat{\mathbf{Y}}$ while preserving the visible chrominance:

$$\hat{\mathbf{R}} = \hat{\mathbf{Y}} + 1.402(\mathbf{Cr}_{\mathrm{v}} - 0.5), \tag{23}$$

$$\hat{\mathbf{G}} = \hat{\mathbf{Y}} - 0.344136(\mathbf{Cb}_{\mathrm{v}} - 0.5) - 0.714136(\mathbf{Cr}_{\mathrm{v}} - 0.5), \tag{24}$$

$$\hat{\mathbf{B}} = \hat{\mathbf{Y}} + 1.772(\mathbf{Cb}_{\mathrm{v}} - 0.5), \tag{25}$$

$$\hat{\mathbf{I}}_{\mathrm{rgb}} = \mathrm{clip}\left((\hat{\mathbf{R}}, \hat{\mathbf{G}}, \hat{\mathbf{B}}), 0, 1\right). \tag{26}$$

Table 10: External assets used in this work. "Not specified" means that no explicit repository-level license is shown on the public repository page.

Asset	Type	Repository / Source	License / Terms
SwinFusion JAS'22	Baseline implementation	https://github.com/Linfeng-Tang/SwinFusion.git	Not specified
TarDAL CVPR'22	Baseline implementation / M3FD source	https://github.com/JinyuanLiu-CV/TarDAL.git	GPL-3.0
CDDFuse CVPR'23	Baseline implementation	https://github.com/Zhaozixiang1228/MMIF-CDDFuse.git	Not specified
Diff-IF Inf. Fusion'24	Baseline implementation	https://github.com/XunpengYi/Diff-IF.git	MIT
EMMA CVPR'24	Baseline implementation	https://github.com/Zhaozixiang1228/MMIF-EMMA.git	Not specified
DCEvo CVPR'25	Baseline implementation	https://github.com/Beate-Suy-Zhang/DCEvo.git	Not specified
Deno-IF NIPS'25	Baseline implementation	https://github.com/hanna-xu/Deno-IF.git	Not specified
SAGE CVPR'25	Baseline implementation	https://github.com/RollingPlain/SAGE_IVIF.git	MIT
UPFusion AAAI'26	Baseline implementation	https://github.com/ixilai/UP-Fusion.git	Not specified
MSRS	Dataset	https://github.com/Linfeng-Tang/MSRS.git	Not specified
FMB	Dataset	https://github.com/JinyuanLiu-CV/SegMiF.git	Not specified
M3FD	Dataset	https://github.com/JinyuanLiu-CV/TarDAL.git	See TarDAL repository; dataset terms not separately specified
IVIF_ZOO metrics	Evaluation code	https://github.com/RollingPlain/IVIF_ZOO	Not specified
SegFormer-B2 Cityscapes checkpoint	Downstream pretrained checkpoint	https://huggingface.co/nvidia/segformer-b2-finetuned-cityscapes-1024-1024/tree/main	Other

Table 11: Complete image-level fusion results on MSRS. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	EN↑	AG↑	MI↑	SD↑	SCD↑	VIF↑	Qabf↑	SSIM↑	SF↑	CE↓
SwinFusion JAS'22	6.620	3.562	4.393	42.980	1.686	1.032	0.647	0.721	11.096	0.052562
TarDAL CVPR'22	6.211	2.926	2.669	32.781	1.375	0.716	0.416	0.540	9.547	0.065997
CDDFuse CVPR'23	6.699	3.749	4.705	43.368	1.618	1.084	0.676	0.709	11.567	0.053406
Diff-IF Inf. Fusion'24	6.669	3.696	**4.970**	42.596	1.624	1.097	0.685	0.711	11.460	0.054479
EMMA CVPR'24	**6.715**	3.786	4.129	**44.568**	1.628	1.016	0.632	0.711	11.574	0.052032
DCEvo CVPR'25	6.622	3.725	4.157	42.926	1.662	1.066	0.688	0.721	11.470	**0.050816**
Deno-IF NIPS'25	6.693	3.112	2.957	41.323	**1.791**	0.902	0.560	0.710	9.727	0.053525
SAGE CVPR'25	6.004	3.156	3.218	36.338	1.417	0.762	0.537	0.689	10.420	0.065907
UPFusion AAAI'26	6.608	3.562	3.618	43.701	1.610	1.009	0.637	**0.725**	11.225	0.051665
PCDFusion Ours	6.711	**3.839**	4.578	42.869	1.671	**1.117**	**0.723**	0.701	**11.632**	0.053841

B Experimental Details

Computational Environment. All training and inference experiments are conducted on a server with one NVIDIA RTX 4090 GPU and two Intel Xeon Platinum 8481C CPUs. Each CPU has 56 physical cores and supports 2 threads per core, resulting in 224 logical CPU threads in total. Under the training setting reported in the main paper, PCDFusion uses 18,206 MiB GPU memory and takes about 18 seconds per epoch on the MSRS training split. For single-image inference on MSRS, the GPU memory usage is about 1,696 MiB.

Datasets and Training Split. PCDFusion is trained only on the MSRS training split, which contains 1,083 aligned visible-infrared image pairs. For image-level fusion evaluation, we use the MSRS test split with 361 pairs, the FMB test split with 280 pairs, and 300 pairs from M3FD. The fusion model is directly applied to all test sets without dataset-specific fine-tuning.

For downstream evaluation, we follow the task setting of each benchmark. For FMB semantic segmentation, we use 1,220 training pairs and 280 testing pairs. For M3FD object detection, the detector is trained on M3FD images outside the 300 fusion-evaluation pairs and evaluated on the fused results of these 300 pairs. Semantic segmentation labels and object detection annotations are used only to train and evaluate downstream task models, not to train PCDFusion.

External Assets and Public Repositories. We use public datasets, official baseline implementations, public metric code, and publicly released downstream checkpoints when available. Tab. 10 lists the external assets used in our experiments. All assets are used only for academic evaluation, following the terms specified by their corresponding repositories.

Baseline and Metric Evaluation. For all compared IVIF methods, we use the official implementations and the publicly released original best checkpoints provided by the authors, without retraining or parameter tuning. Fusion results are generated with the default inference settings of each official codebase on the same test image pairs. We evaluate AG, SCD, VIF, Qabf, and Qcb using the public metric code from IVIF_ZOO. For each dataset, each metric is computed on all test images and averaged over the dataset.

Table 12: Complete image-level fusion results on FMB. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	EN↑	AG↑	MI↑	SD↑	SCD↑	VIF↑	Qabf↑	SSIM↑	SF↑	CE↓
SwinFusion JAS'22	6.683	4.052	4.270	35.352	1.560	1.032	0.665	0.743	13.515	0.000238
TarDAL CVPR'22	**6.963**	3.541	3.423	**40.316**	1.573	0.807	0.462	0.745	11.718	0.000369
CDDFuse CVPR'23	6.778	4.306	4.148	37.043	1.622	1.061	0.672	0.745	14.577	0.000230
Diff-IF Inf. Fusion'24	6.632	4.047	**4.985**	34.191	1.368	1.047	0.660	0.735	13.869	0.000229
EMMA CVPR'24	6.775	**4.684**	3.947	36.805	1.504	1.031	0.645	0.735	**15.006**	0.000224
DCEvo CVPR'25	6.673	4.051	4.512	34.749	1.503	1.058	0.686	0.747	13.771	0.000218
Deno-IF NIPS'25	6.732	3.190	3.004	33.266	1.654	0.830	0.568	**0.767**	11.243	0.000243
SAGE CVPR'25	6.829	3.785	3.429	37.587	**1.709**	0.969	0.637	0.752	12.644	0.000243
UPFusion AAAI'26	6.621	4.097	4.217	33.968	1.427	1.081	0.684	0.747	13.892	0.000212
PCDFusion Ours	6.832	4.334	4.411	38.127	1.568	**1.084**	**0.705**	0.730	14.395	**0.000186**

Table 13: Complete image-level fusion results on M3FD. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	EN↑	AG↑	MI↑	SD↑	SCD↑	VIF↑	Qabf↑	SSIM↑	SF↑	CE↓
SwinFusion JAS'22	6.808	4.606	4.020	35.776	1.560	0.898	0.607	0.717	13.672	0.031855
TarDAL CVPR'22	6.985	4.265	3.130	**38.785**	1.523	0.708	0.427	0.708	12.672	**0.029532**
CDDFuse CVPR'23	6.907	4.861	3.774	37.206	1.646	0.917	0.610	0.718	14.753	0.031766
Diff-IF Inf. Fusion'24	6.791	4.567	**4.852**	34.927	1.322	0.894	0.589	0.701	14.115	0.031833
EMMA CVPR'24	6.927	**5.335**	3.753	38.253	1.494	0.897	0.590	0.703	**15.235**	0.031770
DCEvo CVPR'25	6.842	4.620	4.298	35.582	1.491	0.917	0.633	0.720	14.109	0.031699
Deno-IF NIPS'25	6.789	3.713	2.701	32.569	1.683	0.736	0.528	**0.734**	11.499	0.031226
SAGE CVPR'25	6.850	4.249	3.115	36.465	**1.694**	0.856	0.583	0.724	12.729	0.031960
UPFusion AAAI'26	6.819	4.697	4.023	35.611	1.394	**0.954**	0.637	0.720	14.154	0.031829
PCDFusion Ours	**6.987**	5.037	3.962	38.388	1.575	0.949	**0.668**	0.699	15.216	0.031670

Downstream Evaluation Protocol. For downstream evaluation, semantic and detection labels are used only to train task models, not to train PCDFusion. For FMB semantic segmentation, since FMB uses the same semantic categories as Cityscapes, we initialize SegFormerB2 with a Cityscapes-finetuned checkpoint. For M3FD object detection, we first train a YOLOv10s detector on the M3FD training split using both visible and infrared images for 100 epochs, and use the resulting detector as the pretrained initialization.

To ensure a fair comparison, we generate fused training images for each compared method on the FMB and M3FD training sets. Starting from the same pretrained task model, we then fine-tune the segmentation or detection model on each method's fused training images for 50 epochs. Finally, the fine-tuned task model is evaluated on the corresponding fused test images. All compared fusion methods use the same downstream training data, task labels, pretrained initialization, fine-tuning schedule, and evaluation protocol.

C Additional Quantitative Results

Complete Image-level Fusion Results. Tabs. 11 to 13 provide complete image-level metrics on MSRS, FMB, and M3FD. PCDFusion shows consistent advantages on fidelity and edge-preservation metrics. It achieves the best VIF and Qabf on MSRS and FMB, and obtains the best Qabf with the second-best VIF on M3FD. It also remains strong on gradient- and structure-related metrics, ranking first in AG and SF on MSRS, second in AG and SD on FMB, and second in AG, SD, and SF on M3FD. Although some competing methods lead in individual metrics such as MI, SCD, or SSIM, their advantages are usually dataset- or metric-specific. Overall, the complete results support that PCDFusion provides a stable fusion balance rather than optimizing a single quality indicator.

Table 14: Complete FMB semantic segmentation results. The zero-valued classes are omitted. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	Road	Side.	Build.	Light	Sign	Veg.	Sky	Per.	Car	Truck	Bus	Motor.	Pole	mIoU↑
SwinFusion JAS'22	89.22	48.98	83.16	43.22	80.68	87.16	94.80	70.58	84.23	48.77	57.57	36.28	50.23	58.33
TarDAL CVPR'22	89.50	50.05	83.24	42.30	81.20	88.21	95.00	**73.31**	81.85	35.82	26.36	37.18	49.76	55.59
CDDFuse CVPR'23	**90.18**	50.97	84.38	40.85	81.80	**88.40**	95.16	72.22	82.90	22.14	55.70	33.60	51.56	56.66
Diff-IF Inf. Fusion'24	89.66	43.81	83.64	45.24	82.04	87.02	94.70	70.69	83.57	45.22	**75.99**	32.26	50.82	58.98
EMMA CVPR'24	89.61	49.32	**84.57**	**49.42**	81.24	88.15	**95.29**	70.75	82.58	42.99	59.41	32.61	51.07	58.47
DCEvo CVPR'25	89.72	48.01	84.39	46.98	81.01	88.14	95.27	71.50	**84.55**	43.67	49.08	33.89	49.73	57.73
Deno-IF NIPS'25	88.89	47.39	84.05	40.84	80.57	88.34	95.01	72.73	82.25	40.13	34.40	29.23	47.86	55.45
SAGE CVPR'25	89.97	**51.65**	83.45	42.90	**82.37**	88.08	94.90	71.45	82.28	41.08	31.51	33.66	**51.61**	56.33
UPFusion AAAI'26	88.83	43.86	84.23	19.68	77.97	87.64	95.00	69.73	81.92	39.27	58.29	34.85	49.21	55.36
PCDFusion Ours	89.37	50.79	84.36	45.89	81.31	88.00	95.07	70.23	84.44	**51.54**	71.19	**37.98**	50.60	**60.05**

Table 15: Complete M3FD object detection results. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	mAP_{50}							$mAP_{50:95}$						
	Per.	Car	Bus	Motor.	Lamp	Truck	mAP	Per.	Car	Bus	Motor.	Lamp	Truck	mAP
SwinFusion JAS'22	82.28	91.25	89.10	66.22	79.83	78.75	81.24	48.90	68.13	71.31	40.82	42.03	58.00	54.86
TarDAL CVPR'22	81.94	90.88	87.43	59.35	78.67	76.25	79.09	48.80	66.51	69.55	34.00	38.87	56.38	52.35
CDDFuse CVPR'23	81.93	91.57	90.26	68.36	81.59	74.18	81.31	48.79	68.44	71.80	38.00	43.44	54.39	54.15
Diff-IF Inf. Fusion'24	82.81	91.38	88.46	66.32	77.91	79.14	81.00	49.77	67.43	70.08	38.47	39.82	58.68	54.04
EMMA CVPR'24	81.70	91.11	**90.46**	63.35	81.59	77.36	80.93	48.35	68.03	71.02	38.26	43.70	57.20	54.43
DCEvo CVPR'25	**83.59**	91.52	88.96	**71.23**	78.56	**80.10**	**82.33**	**50.05**	68.25	71.60	41.20	43.86	**59.78**	55.79
Deno-IF NIPS'25	83.34	91.88	88.38	69.87	77.82	78.98	81.71	50.03	68.68	70.37	41.06	43.16	57.61	55.15
SAGE CVPR'25	82.34	91.30	86.46	65.82	78.10	75.46	79.91	49.59	68.42	68.91	39.15	43.22	56.39	54.28
UPFusion AAAI'26	82.09	91.82	87.90	65.47	**82.29**	76.73	81.05	48.72	68.52	71.28	38.57	**45.94**	58.04	55.18
PCDFusion Ours	82.30	**91.93**	90.04	69.53	79.38	76.99	81.69	48.95	**69.00**	**72.60**	**42.89**	44.53	57.93	**55.98**

Complete downstream results. Tabs. 14 and 15 report complete downstream results on FMB semantic segmentation and M3FD object detection. For segmentation, zero-valued classes are omitted from the table for readability, while the reported mIoU follows the same evaluation setting as the main paper. PCDFusion achieves the best overall mIoU and performs strongly on object-related and thin-structure categories, including Car, Truck, Bus, and Motorcycle. For detection, PCDFusion obtains the best $mAP_{50:95}$ and the second-best mAP_{50}, with clear gains on Car, Bus, and Motorcycle under the stricter $mAP_{50:95}$ metric. These results indicate that the fused images produced by PCDFusion support both semantic structure parsing and object localization, rather than improving only image-level fusion scores.

Local structure recovery and appearance preservation. The first MSRS example in Fig. 3 contains a locally saturated building whose contour remains visible in the infrared observation. To separate useful structure recovery from indiscriminate infrared transfer, we measure infrared-supported edge retention on over-exposed visible pixels with strong infrared gradients and reliable-sky preservation on sky pixels whose fused luminance remains within 0.05 of the visible input. As shown in Tab. 16, PCDFusion achieves the highest edge retention, 66.3%, while preserving 99.5% of reliable-sky pixels. This combination indicates that the recovered contour is localized to the failed visible region rather than produced by globally copying infrared edges.

Single-modality-to-fused transfer. We additionally evaluate zero-shot compatibility with task models trained on individual modalities. Visible-only and infrared-only task models are trained separately and then evaluated directly on each method's fused test images; the reported mean is the arithmetic mean of the two scores. All methods use the same data splits, task architectures, and evaluation code, and each entry is averaged over three task-model seeds. Tab. 17 shows that PCDFusion ranks first on both aggregate means, reaching 48.77% mIoU on FMB and 19.80% $mAP_{50:95}$ on M3FD. The result complements the task-adapted protocol in the main paper by showing that the fused images also transfer well to models that have not been trained on fused data.

Table 16: Local structure recovery and appearance preservation in the first MSRS visualization. The best result is highlighted in **bold**, and the second-best result is underlined.

Metric	SwinFusion	TarDAL	CDDFuse	Diff-IF	EMMA	DCEvo	Deno-IF	SAGE	UPFusion	PCDFusion
Over-exposed IR-edge retention (%)↑	16.8	27.0	51.3	47.4	24.0	62.9	49.0	25.7	17.3	**66.3**
Reliable-sky preservation (%)↑	**100.0**	0.1	99.5	99.6	97.5	99.5	7.3	0.0	99.9	99.5

Table 17: Single-modality-to-fused transfer results. VIS-trained and IR-trained task models are evaluated directly on fused test images; Mean averages the two scores. The best result is highlighted in **bold**, and the second-best result is underlined.

Fusion method	FMB mIoU (%)			M3FD $mAP_{50:95}$ (%)		
	VIS-trained	IR-trained	Mean	VIS-trained	IR-trained	Mean
SwinFusion	54.18	40.39	47.28	25.96	8.46	17.21
TarDAL	51.76	**42.46**	47.11	24.48	10.04	17.26
CDDFuse	56.09	40.79	48.44	26.54	8.35	17.44
Diff-IF	53.61	40.01	46.81	23.30	10.34	16.82
EMMA	55.65	39.59	47.62	30.01	5.93	17.97
DCEvo	54.96	40.84	47.90	26.12	9.32	17.72
Deno-IF	54.28	41.23	47.75	26.57	**10.91**	18.74
SAGE	54.76	41.45	48.10	26.99	9.89	18.44
UPFusion	55.96	40.13	48.04	32.28	6.71	19.50
PCDFusion	**56.36**	41.18	**48.77**	**32.30**	7.31	**19.80**

Comparison with feature-level RGB-IR detection. Pixel-level fusion produces a reusable image for visualization and conventional single-image task models, whereas feature-level detectors consume paired visible and infrared inputs through a task-specific architecture. We compare these two paradigms on the same 300-image M3FD test set using the same evaluation code and closely matched model capacity. As reported in Tab. 18, PCDFusion with YOLOv10x reaches 61.05% $mAP_{50:95}$, only 0.22 points below M2D-LIF and 1.67 points above LoME. Thus, PCDFusion approaches the strongest feature-level result while retaining a viewable fused output and single-image inference for downstream deployment.

D Additional Qualitative Results

More Image-level fusion visualizations. Fig. 7 provides additional qualitative comparisons on challenging scenes. In the first scene, the visible image contains an over-exposed building region, while the infrared image preserves its coarse structure. Most competing methods either keep the region saturated or recover only weak contours. PCDFusion better restores the building structure while maintaining the surrounding visible appearance. In the second scene, the infrared image clearly highlights the street lamp under low-contrast visible observation. PCDFusion effectively incorporates the lamp response without spreading excessive thermal intensity to the background.

More downstream visualizations. Fig. 8 shows additional downstream results. For semantic segmentation, PCDFusion produces cleaner masks and almost completely segments the vegetation and pedestrians in the scene. For object detection, PCDFusion accurately detects the motorcycle and the rider on it, while avoiding obvious background interference. These results further indicate that the fused images preserve task-relevant structures and targets.

E Additional Discussions

Additional Prior Variants. The main paper reports the ablation of posterior estimation forms and context prior estimates. Here, we further provide the ablation of proposal and detail prior variants. These variants test whether the stage-specific physical priors can be replaced by simpler estimates. We denote the normalized visible gradient by $\mathbf{S}_v = \mathrm{Norm}(|\nabla \mathbf{Y}_v|)$, the normalized infrared edge

Table 18: Capacity-matched comparison with feature-level RGB-IR detectors on M3FD. Parameter counts include the complete training architectures, and GFLOPs are measured for deployed 640×640 inference.

Method	Fusion level	Test-time inputs	Training protocol	Params (M)↓	GFLOPs↓	$mAP_{50:95}$ (%)↑
PCDFusion + YOLOv10x	Pixel/image	One fused image	300 epochs	36.27	508.62	61.05
M2D-LIF	Feature	Paired VIS+IR	200-epoch teacher + 100-epoch student	36.51	122.42	**61.26**
LoME (width 0.90)	Feature	Paired VIS+IR	300 epochs	36.18	49.05	59.38



Figure 7: Additional visual comparisons on MSRS and FMB.

response by $\mathbf{S}_{er} = \mathrm{Norm}(|\nabla \mathbf{Y}_r|)$, and the combined infrared structure by $\mathbf{S}_r = \max(\mathbf{S}_{er}, \mathbf{S}_{br})$, where $\mathbf{S}_{br}$ is the band-limited infrared response. For the proposal stage, constant or hard priors can slightly increase AG or SCD, but they reduce VIF, Qcb, mIoU, and mAP compared with the full model. This suggests that the soft thermal-response prior is more suitable for preserving possible thermal responses without over-constraining later rendering. For the detail stage, edge-, structure-, or intensity-based variants improve some difference-related scores, but the full model achieves stronger fidelity and downstream performance. This supports the use of structure-competitive detail prior together with posterior calibration, rather than directly rendering infrared edges or intensities.

Fixed-checkpoint parameter sensitivity. To isolate inference-time sensitivity from retraining variation, we keep the Full checkpoint fixed and vary one scalar at a time. Tab. 20 reports the absolute values together with their changes from the default. Across proposal temperatures from 2 to 10, FMB mIoU and M3FD mAP vary by only 0.12 and 0.11 points, respectively. Across detail scales from 0.05 to 0.15, the corresponding spans are 0.22 and 0.29 points. The stable downstream performance over these broad ranges indicates that the gains do not depend on a narrowly tuned inference scalar.

Stage and objective trade-offs. The metric reversals in the stage ablation follow from the distinct image properties modified by the three renderers. Proposal rendering maintains high-recall thermal candidates; removing it discards target evidence before later selection and lowers mIoU and mAP by 1.61 and 0.80 points. Context rendering repairs exposure-failed visible regions and suppresses unsupported variation. Without it, SCD increases from 1.671 to 1.716 because more cross-modal contrast remains, but mIoU and mAP decrease by 1.19 and 1.10 points. Detail rendering injects a signed band-limited residual under structure competition, improving VIF, Qcb, and mIoU while allowing a small decrease in AG. The 0.15-point mAP advantage without this stage reflects that detector confidence is not optimized directly and need not vary monotonically with all fusion-quality measures. The Full model therefore provides the strongest overall balance rather than maximizing every scalar independently.

The objective ablation exhibits a related hierarchy. $\mathcal{L}_d$ supervises the final fused luminance densely and directly, whereas $\mathcal{L}_p$, $\mathcal{L}_c$, and the keep term constrain intermediate proposal and context decisions.

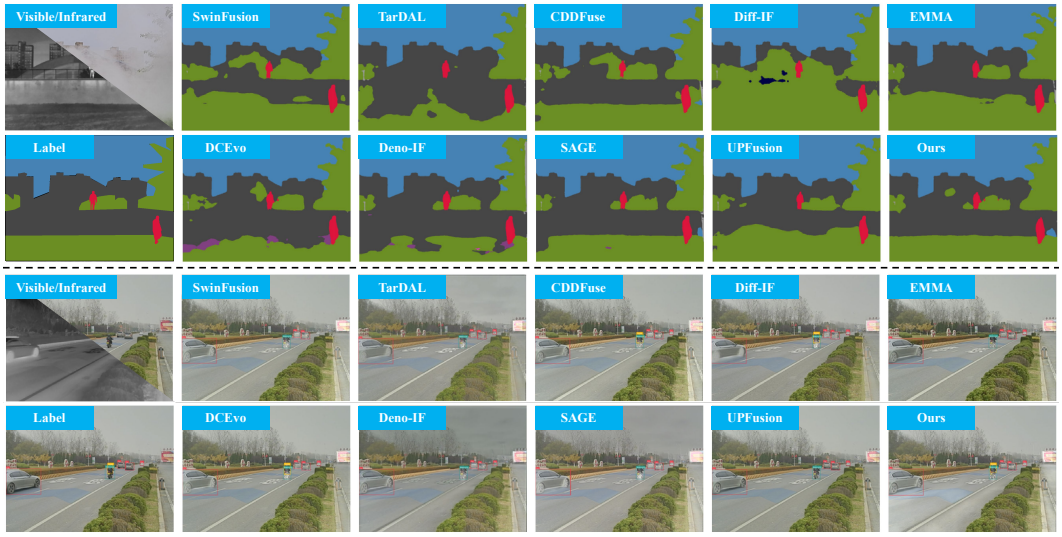


Figure 8: Additional FMB semantic segmentation and M3FD object detection visualizations.

Table 19: Additional ablation of proposal and detail prior variants. The best results are highlighted in **bold**, and the second-best results are underlined.

Stage	Variant	AG↑	SCD↑	VIF↑	Qabf↑	Qcb↑	mIoU↑	mAP↑
Proposal	Uniform: 0.5	**3.863**	**1.672**	1.110	**0.723**	0.591	58.88	55.86
Proposal	Hard IR: $\mathbb{1}[\mathrm{Norm}(\mathbf{Y}_r) \geq 0.5]$	3.861	1.671	1.109	0.722	0.589	58.95	55.72
Proposal	Soft IR (Full): $\sigma(\tau_p(\mathrm{Norm}(\mathbf{Y}_r) - 0.5))$	3.839	1.671	**1.117**	**0.723**	**0.593**	**60.05**	**55.98**
Detail	Edge competition: $\mathbf{S}_{er}/(\mathbf{S}_{er} + \mathbf{S}_v + \epsilon)$	3.855	1.674	1.104	0.721	0.585	58.95	55.83
Detail	IR structure only: $\mathbf{S}_r$	3.848	1.677	1.110	0.722	0.587	58.84	55.70
Detail	IR intensity: $\mathrm{Norm}(\mathbf{Y}_r)$	**3.859**	**1.678**	1.105	0.721	0.586	58.90	55.95
Detail	Structure competition (Full): $\mathbf{S}_r/(\mathbf{S}_r + \mathbf{S}_v + \epsilon)$	3.839	1.671	**1.117**	**0.723**	**0.593**	**60.05**	**55.98**

When one intermediate constraint is removed, later renderers and $\mathcal{L}_d$ can compensate for part of the missing guidance, producing smaller changes in aggregate downstream scores. Nevertheless, these terms prevent different failure modes: $\mathcal{L}_p$ preserves infrared-salient candidates, $\mathcal{L}_c$ anchors context updates to visible failure with infrared support, and the keep term suppresses unnecessary updates in reliable visible regions. Their joint use preserves the intended proposal-context-detail behavior across fusion and downstream criteria.

Posterior-action interventions. We directly test whether the spatial values of the posterior maps contribute to local rendering quality. With the checkpoint and inputs fixed, one learned posterior-action map at a time is replaced by its prior, a spatially shuffled map, an inverted map, or a zero map. As shown in Tab. 21, the learned posterior provides the largest local task gain at every stage and the best final FMB mIoU in every intervention group. Shuffling the proposal or context map reduces mIoU from 60.05 to 50.21 or 50.16, respectively, demonstrating that the observation-conditioned spatial assignment, rather than only the mean weight magnitude, is important for the fused result.

Computational cost. Tab. 22 reports the model size, GFLOPs, inference time, and FPS under the same 512×512 input setting. PCDFusion has 4.60M parameters and requires 235.34 GFLOPs, which is higher than several lightweight baselines but much lower than large generative or heavy interaction-based methods such as Diff-IF and UPFusion. Its inference speed is also substantially faster than SwinFusion, Diff-IF, and UPFusion. This indicates that PCDFusion achieves the reported fusion and downstream gains with moderate computational cost, although it is not designed as the fastest lightweight model.

Exposure-Failure Prior Calibration. We further examine whether the analytic exposure-failure prior agrees with empirical visible failures. This analysis is only used for post-hoc validation and is

Table 20: Fixed-checkpoint parameter sensitivity. Parenthesized values show changes from the default setting.

Parameter	Value	AG	VIF	Qabf	mIoU	mAP
Proposal temperature	2	3.813 (−0.026)	1.115 (−0.002)	0.720 (−0.003)	60.11 (+0.06)	55.94 (−0.05)
	4	3.832 (−0.007)	1.119 (+0.002)	0.723 (0.000)	60.10 (+0.05)	55.88 (−0.11)
	6 (default)	3.839 (0.000)	1.117 (0.000)	0.723 (0.000)	60.05 (0.00)	55.98 (0.00)
	8	3.830 (−0.009)	1.104 (−0.013)	0.720 (−0.002)	60.03 (−0.02)	55.90 (−0.08)
	10	3.794 (−0.045)	1.080 (−0.037)	0.714 (−0.009)	59.99 (−0.06)	55.95 (−0.04)
Detail scale	0.050	3.729 (−0.110)	1.109 (−0.008)	0.714 (−0.008)	59.90 (−0.15)	55.88 (−0.10)
	0.075	3.782 (−0.057)	1.114 (−0.003)	0.720 (−0.003)	60.00 (−0.05)	55.84 (−0.14)
	0.100 (default)	3.839 (0.000)	1.117 (0.000)	0.723 (0.000)	60.05 (0.00)	55.98 (0.00)
	0.125	3.898 (+0.060)	1.118 (+0.001)	0.722 (−0.001)	60.11 (+0.06)	56.00 (+0.02)
	0.150	3.960 (+0.122)	1.117 (0.000)	0.720 (−0.003)	60.12 (+0.07)	55.72 (−0.27)

Table 21: Posterior-action interventions and local utility. Each intervention replaces one stage map while keeping the checkpoint and inputs fixed. Local gains and FMB mIoU are reported in percentage points.

Stage	Posterior-action map used for rendering	Local task gain↑	FMB mIoU↑
Proposal	Learned posterior	**0.071**	**60.05**
	Prior only	0.069	59.86
	Spatially shuffled	−0.491	50.21
	Inverted	0.003	58.68
	Zero map	0.000	58.58
Context	Learned posterior	**0.013**	**60.05**
	Prior only	0.003	59.36
	Spatially shuffled	−0.519	50.16
	Inverted	−0.020	57.66
	Zero map	0.000	58.68
Detail	Learned posterior	**0.0029**	**60.05**
	Prior only	0.0014	59.97
	Spatially shuffled	0.0018	59.73
	Inverted	−0.0004	59.42
	Zero map	0.0000	59.54

not used to train PCDFusion. We use a semantic segmentation model fine-tuned on MSRS to estimate how often visible observations fail under different luminance levels.

For each visible image, let $\mathbf{Y}_v$ be the visible luminance, $\hat{\mathbf{M}}_v$ be the semantic prediction from the visible image, and $\mathbf{M}$ be the ground-truth semantic label. The visible-failure map is defined as

$$\mathbf{E}_v(x) = \mathbb{1}\left[\hat{\mathbf{M}}_v(x) \neq \mathbf{M}(x)\right], \tag{27}$$

where x denotes a pixel. We divide visible luminance into B bins. For bin b, the pixel-level empirical failure is computed as

$$F_b^{\mathrm{pix}} = \frac{\sum_{x\in\Omega_b} \mathbf{E}_v(x)}{|\Omega_b| + \epsilon}, \qquad \Omega_b = \{x \mid \mathbf{Y}_v(x) \in \mathcal{B}_b\}. \tag{28}$$

To reduce the bias from dominant semantic categories, we also compute a class-normalized failure score:

$$F_b^{\mathrm{cls}} = \frac{1}{|\mathcal{C}_b|} \sum_{c\in\mathcal{C}_b} \frac{\sum_{x\in\Omega_b^c} \mathbf{E}_v(x)}{|\Omega_b^c| + \epsilon}, \qquad \Omega_b^c = \{x \mid \mathbf{Y}_v(x) \in \mathcal{B}_b,\ \mathbf{M}(x) = c\}, \tag{29}$$

Table 22: Computational cost comparison on 512×512 inputs. The best results are highlighted in **bold**, and the second-best results are underlined.

Metric	SwinFusion IAS'22	TarDAL CVPR'22	CDDFuse CVPR'23	Diff-IF Inf. Fusion'24	EMMA CVPR'24	DCEvo CVPR'25	Deno-IF NIPS'25	SAGE CVPR'25	UPFusion AAAI'26	PCDFusion Ours
Params (M)↓	0.97	0.30	1.19	23.74	1.52	2.00	1.43	**0.14**	154.82	4.60
GFLOPs↓	604.42	155.07	945.04	6101.12	68.46	1561.87	221.15	**34.11**	2676.36	235.34
Time (ms)↓	997.82	5.04	126.70	780.17	13.17	178.92	112.68	**2.12**	584.92	92.51
FPS↑	1.00	198.55	7.89	1.28	75.92	5.59	8.87	**471.34**	1.71	10.81

Table 23: Calibration of exposure-failure estimates on MSRS and FMB. The best results are highlighted in **bold**, and the second-best results are underlined.

Dataset	Estimate	Pixel failure			Class-normalized mFailure		
		Pearson↑	Spearman↑	MSE↓	Pearson↑	Spearman↑	MSE↓
MSRS	Darkness-only: $1-\mathbf{Y}_v$	**0.762**	**0.982**	0.1372	-0.111	-0.035	0.2231
	Brightness-only: $\mathbf{Y}_v$	-0.762	-0.982	0.3381	0.111	0.035	0.1936
	Distance-to-mid: $\lvert 2\mathbf{Y}_v-1\rvert$	0.418	0.087	0.1825	0.815	**0.928**	0.0998
	Ours prior: $(2\mathbf{Y}_v-1)^2$	0.463	0.087	0.1014	0.879	**0.928**	**0.0314**
	Ours posterior: $\mathbf{q}_c$	0.646	0.147	**0.0552**	**0.890**	0.904	0.0658
FMB	Darkness-only: $1-\mathbf{Y}_v$	0.524	0.432	0.1286	**0.563**	**0.555**	**0.0691**
	Brightness-only: $\mathbf{Y}_v$	-0.524	-0.432	0.2849	-0.563	-0.555	0.2393
	Distance-to-mid: $\lvert 2\mathbf{Y}_v-1\rvert$	0.530	0.490	0.1261	0.369	0.405	0.0983
	Ours prior: $(2\mathbf{Y}_v-1)^2$	0.565	0.490	0.0722	0.445	0.405	0.1358
	Ours posterior: $\mathbf{q}_c$	**0.661**	**0.737**	**0.0412**	0.507	0.435	0.1733

where $\mathcal{C}_b$ is the set of semantic classes appearing in bin b. Both empirical failure curves and candidate exposure estimates are normalized to $[0,1]$ before computing MSE.

For each bin, we compare the empirical failure score with the mean analytic prior

$$U_b = \frac{1}{|\Omega_b| + \epsilon} \sum_{x \in \Omega_b} (2\mathbf{Y}_v(x) - 1)^2. \tag{30}$$

We report Pearson correlation, Spearman correlation, and MSE. Given two bin-wise sequences $\mathbf{a} = \{a_b\}_{b=1}^B$ and $\mathbf{f} = \{f_b\}_{b=1}^B$, Pearson correlation is

$$\rho_P = \frac{\sum_{b=1}^B (a_b - \bar{a})(f_b - \bar{f})}{\sqrt{\sum_{b=1}^B (a_b - \bar{a})^2} \sqrt{\sum_{b=1}^B (f_b - \bar{f})^2}}. \tag{31}$$

Spearman correlation is Pearson correlation computed on ranks:

$$\rho_S = \rho_P \left(\mathrm{rank}(\mathbf{a}), \mathrm{rank}(\mathbf{f}) \right). \tag{32}$$

The MSE is computed as

$$\mathrm{MSE} = \frac{1}{B} \sum_{b=1}^B (a_b - f_b)^2. \tag{33}$$

Tab. 23 and Fig. 9 analyze the relation between exposure estimates and empirical visible failure. The two empirical statistics measure different aspects. Pixel failure is dominated by frequent pixels and large regions, so it mainly reflects where the visible image actually produces dense prediction errors. Class-normalized mFailure gives each semantic class a more balanced contribution within each luminance bin, making it more sensitive to rare classes and category composition. Therefore, a prior or posterior can fit one statistic better than the other.

On MSRS, the analytic exposure prior $U_v = (2\mathbf{Y}_v - 1)^2$ agrees well with class-normalized mFailure. It achieves strong class-level Pearson and Spearman correlations and the lowest class-level MSE among analytic estimates. This supports the basic physical assumption that both darkness and saturation increase visible unreliability. However, the learned posterior $\mathbf{q}_c$ is not optimized to

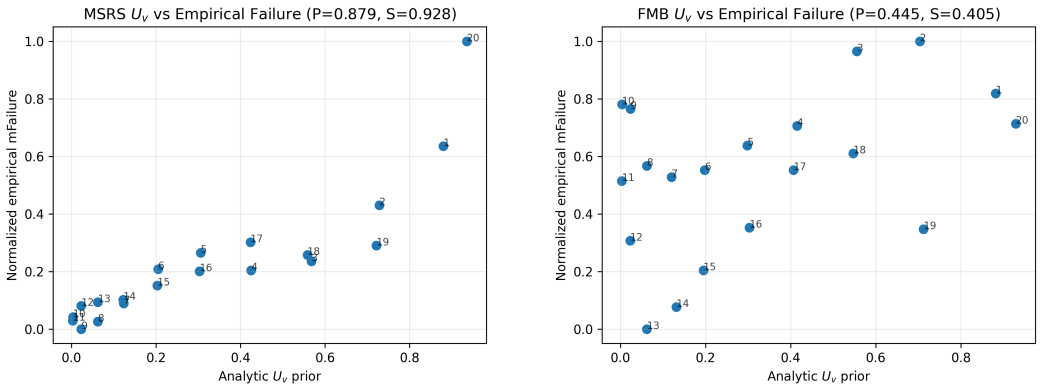


Figure 9: Relation between the analytic exposure-failure prior U_v and empirical class-normalized visible failure on MSRS and FMB. Each point denotes a luminance bin, and Pearson/Spearman correlations are reported in the subplot titles.

reproduce this empirical curve. It is a rendering weight calibrated by cross-modal observations. As a result, $\mathbf{q}_c$ improves pixel-level fitting, especially in MSE, but can be weaker than the raw prior in class-normalized MSE or rank correlation. This indicates that the posterior does not simply preserve the exposure prior; it suppresses updates in some exposure-failure regions when infrared structure or local observations do not support context recovery.

The difference is more evident on FMB. The class-normalized mFailure on FMB is affected by dataset composition and semantic-category distribution, and the segmentation model used for this post-hoc analysis is not used to train PCDFusion. In this setting, darkness-only estimation can correlate better with class-normalized mFailure, while $\mathbf{q}_c$ is worse on class-balanced MSE. This does not contradict the role of the posterior. The context posterior is designed to decide where context rendering should be applied, not to be a class-balanced semantic failure predictor. Consistently, $\mathbf{q}_c$ achieves the best pixel-level Pearson, Spearman, and MSE on FMB, showing that it better matches dense visible-failure regions relevant to image rendering. Overall, the analytic prior provides an interpretable exposure-failure trend, while the posterior further calibrates this trend according to local cross-modal support. The calibration is selective rather than uniformly stronger, which is consistent with the goal of avoiding unnecessary context updates in reliable visible regions.

F Failure Cases and Limitations

Fig. 10 provides a closer look at the remaining limitations of PCDFusion. The first scene shows a typical advantage of our posterior rendering design. The visible building region is severely over-exposed, while the infrared image still contains coarse structural cues. Conservative methods such as EMMA and UPFusion preserve the visible appearance more strictly, but they also keep the saturated building region largely unresolved. In contrast, PCDFusion renders more infrared-supported structure in the failed visible area. This result is consistent with the goal of the proposal and context stages, which keep positive thermal responses available and use them to revise exposure-degraded regions.

The second and third scenes reveal the opposite side of this design. In these cases, the visible image still contains useful weak structures, while the infrared observation is less reliable for the local appearance. Some methods, especially EMMA and UPFusion, behave more conservatively and better preserve the original visible structures with less thermal transfer. PCDFusion suppresses obvious infrared contamination more effectively than TarDAL, Diff-IF, and DCEvo, but it still introduces mild thermal interference in the low-luminance regions. The effect is visible around weak branches, dark background regions, and low-contrast structures, where infrared responses can be partially rendered even when they do not correspond to reliable visible appearance.

This limitation is related to the role of the context posterior action weight. PCDFusion is designed to recover visible failure, so its posterior tends to remain responsive in dark or low-contrast regions. As shown in Fig. 6, the context posterior is relatively high in low-luminance ranges. This behavior helps

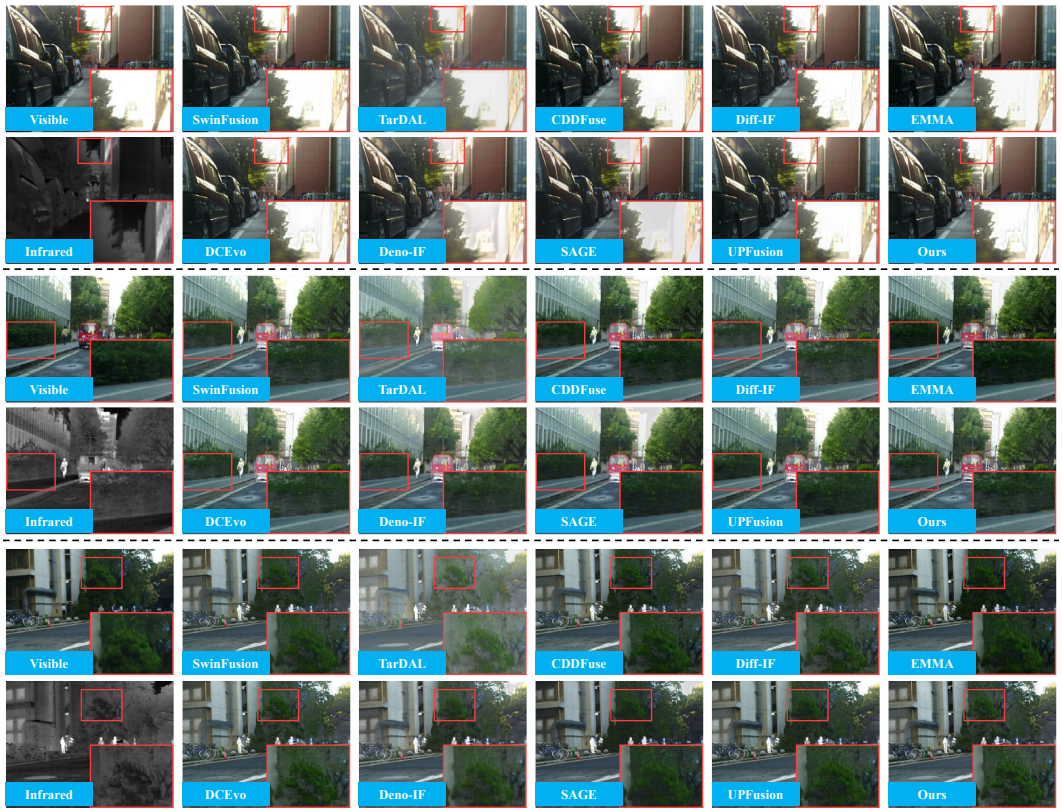


Figure 10: Failure cases and limitations.

recover genuinely degraded visible regions, but it can also make the model less conservative when the infrared observation is unreliable, weakly aligned, or dominated by smooth thermal background. Therefore, the current model still has a trade-off between recovering missing structures and preserving weak but valid visible appearance.

Future work should improve this trade-off in two directions. First, the context posterior can incorporate stronger visible-structure preservation, so that low-luminance updates are reduced when visible edges or thin structures remain consistent. Second, infrared reliability estimation should explicitly detect cases where thermal responses are likely invalid, such as glass reflection, smooth thermal background, and weakly structured low-light regions. A more adaptive posterior calibration may allow the model to keep its advantage in high-exposure recovery while further reducing residual thermal interference in reliable visible regions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main formulation, the PCDFusion architecture, the absence of semantic labels for fusion training, and the claimed improvements on image-level and downstream tasks. These claims are supported by the method description, quantitative comparisons, ablation studies, and limitations discussion.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The main paper includes a limitations paragraph in the discussion, and the supplemental material provides a separate failure-case analysis. The discussion notes remaining issues in low-luminance and infrared-unreliable regions and relates them to posterior calibration.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not claim new theoretical theorems or formal guarantees. It introduces a Bayesian posterior rendering formulation and uses it as a modeling principle for the proposed architecture.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper reports datasets, splits, training settings, loss functions, model architecture, downstream evaluation protocol, baseline usage, metric implementation, and RGB reconstruction details in the main paper and supplemental material. Code is planned for public release upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use public datasets and public metric code, and the implementation code has been released at https://github.com/BIT-RuiLiu/PCDFusion.git.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main paper and supplemental material specify the MSRS, FMB, and M3FD splits, training schedule, optimizer, learning rate, batch size, crop size, downstream fine-tuning details, baseline checkpoint usage, and metric computation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports deterministic dataset-level averages over standard test splits but does not include repeated runs, confidence intervals, or statistical significance tests. This choice follows common IVIF evaluation practice, but the absence of error bars is a limitation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The supplemental material reports the GPU type, CPU configuration, training GPU memory, per-epoch training time, and single-image inference memory. It also reports model parameters, GFLOPs, inference time, and FPS under a fixed input size.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public benchmark datasets and does not involve private data collection, human-subject experiments, or manipulative user-facing systems. To the best of our knowledge, the research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper focuses on the technical formulation and evaluation of IVIF and does not include a dedicated broader-impact discussion. Potential benefits include improved perception under challenging illumination, while possible risks include use in surveillance or safety-critical systems without sufficient validation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk pretrained generative model, language model, or scraped dataset. It uses existing public IVIF benchmarks and proposes an image fusion model.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The supplemental material lists the public repositories or sources of all datasets, baseline implementations, metric code, and downstream pretrained checkpoints used in the experiments. It also reports the visible repository-level license or terms when available, and marks assets as not specified when no explicit license is shown on the public repository page.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce a new dataset or benchmark. The implementation code is planned for release upon acceptance, but no new asset package is provided with the current anonymous submission.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or research with human participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the proposed method, experiments, or evaluation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.