MODULE: A Mutual-Promoting Deep Unfolding Framework Towards Degradation-Robust Multi-modal Image Fusion

Yunfei Deng[1], **Han Xu**[1]*, **Jiayi Ma**[2], **Guangcan Liu**[1]
[1]School of Automation, Southeast University, Nanjing, China
[2]School of Robotics, Wuhan University, Wuhan, China
xu_han@seu.edu.cn, {yunfeideng455, jyma2010}@gmail.com,
guangcanliu@seu.edu.cn

Abstract

Multi-modal image fusion is crucial for comprehensive scene representation, yet real-world degradations compromise its efficacy, necessitating degradation-robust fusion paradigms. Existing degradation-robust multi-modal image fusion methods are often hindered by cascaded sub-optimal solution or black-box opacity. To address these challenges, we propose a theory-inspired mutual-promoting deep unfolding framework. It reformulates degradation-robust image fusion as a joint optimization problem, utilizing the degradation model and modality generation mechanism to explicitly model both single-task priors and cross-task dependencies. By decomposing the complex multi-task optimization problem into task-related iterative subproblems, the framework establishes a bidirectional reciprocal information flow between restoration and fusion. This process is further unfolded into a multi-stage deep neural network, where each component explicitly corresponds to a specific mathematical operation. In our framework, the fused cross-modal prior actively regularizes the ill-posed restoration, while the purified modality features continuously refine the fusion output. It ensures a transparent architecture that combines the merits of both model-based and data-driven methodologies. Extensive experiments demonstrate that our paradigm achieves state-of-the-art performance while providing superior interpretability. The code is publicly available at https://github.com/Meskend/MODULE.

1 Introduction

Due to complex imaging environments, single-modal images often fail to deliver a comprehensive representation of scenes. While visible images offer rich textures, they are compromised in low-light or occluded conditions. Conversely, infrared images can highlight thermal targets but lack fine-grained details. To bridge this gap, multi-modal image fusion (MMIF) aims to integrate complementary information to synthesize a single comprehensive representation. It is widely applied in diverse domains, *e.g.*, autonomous driving [9], intelligent security [23], and military surveillance [10].

In recent years, learning-based multi-modal image fusion (MMIF) methods have achieved remarkable advancements, which can be categorized into methods based on CNN [18, 19], GAN [14, 15], Transformer [22, 25], Mamba [26, 12], and Diffusion [39, 31, 27]. However, they are fundamentally tailored to fusion, inherently assuming that source images are perfectly clean, overlooking complex degradations (*e.g.*, noise, blur, and low light) that frequently occur under challenging imaging

*Corresponding author.

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

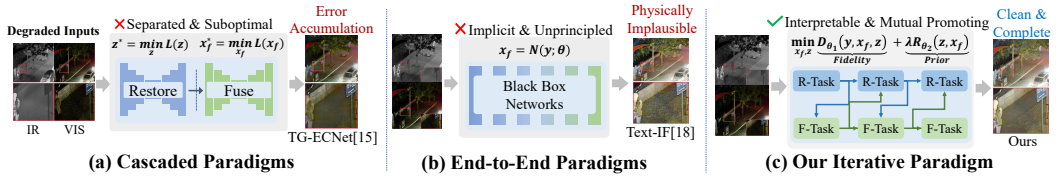


Figure 1: Comparison of existing degradation-robust image fusion paradigms. (a) Cascade paradigms suffer from biased task prior, causing error accumulation. (b) End-to-end paradigms abort the task-specific prior, leading to physically implausible results. (c) Our iterative paradigm is guided by task prior in optimization objectives, resulting in clean and comprehensive results.

conditions. Consequently, their performance deteriorates significantly when deployed in practical scenarios, underscoring the urgent need for degradation-robust multi-modal image fusion.

Fundamentally, degradation-robust image fusion is not a sequential combination of isolated tasks, as restoration and fusion share a profound intrinsic dependency. High-quality restoration provides reliable features for fusion, while the fused output offers cross-modal guidance regularizing ill-posed restoration. As a result, the clean scene representation z and the final fused result x_f should be simultaneously estimated. This inherent connection naturally abstracts the degradation-robust fusion task into a joint optimization problem:

$$\min_{z,x_f} \mathcal{L}(z, x_f) = \mathcal{L}_{res}(z) + \mathcal{L}_{fus}(x_f) + \mathcal{L}_{couple}(z, x_f), \tag{1}$$

where $\mathcal{L}_{res}(z)$ and $\mathcal{L}_{fus}(x_f)$ govern the single-task priors, and $\mathcal{L}_{couple}(z, x_f)$ explicitly encodes this intrinsic task dependency as a cross-task prior.

To tackle this challenging degradation-robust image fusion task, recent pioneering works have emerged, which can be broadly categorized into two main paradigms: cascaded methods and end-to-end black-box methods. However, when evaluated against the aforementioned theoretical yardstick, both paradigms exhibit inherent limitations when encoding task priors into their architecture design.

Cascaded methods, whether concatenating independent algorithms or decoupling processes sequentially within a single network [20, 35, 17, 29], inherently adopt a biased prior encoding strategy, as Fig. 1 (a) shows. These paradigms can be interpreted as solving the joint objective through a one-step sequential decomposition. By simply minimizing the single-task restoration term $\mathcal{L}_{res}(z)$ and completely ignoring the coupling term $\mathcal{L}_{couple}(z, x_f)$ at the initial stage, their final solution (z^*, x_f^*) frequently leaves a non-zero gradient for the coupling term (*i.e.*, $\nabla_z \mathcal{L}_{couple}(z^*, x_f^*) \neq 0$). This implies that cascaded strategies violate the mathematical optimality condition, trapping the solution in a suboptimal state. This mathematical suboptimality translates directly into a biased structural prior within the network architecture. By hard-coding a unidirectional information flow, the architecture forces the model to learn a biased mapping that prohibits exploring better solutions in the joint space. Consequently, this causes irreversible error accumulation: any degradation artifacts or over-smoothed details generated initially cannot be rectified in the subsequent fusion stage.

End-to-end black-box methods [32, 28] completely abandon the explicit modeling of the joint constraints and coupling rules formulated above, as Fig. 1 (b) shows. These task-prior-agnostic architectures favor the heuristic stacking of generic operators such as convolutions or Transformers, instantiating neither a specific degradation model nor a dedicated coupling mechanism. With task dependencies left to be implicitly inferred purely from training data, the coupled processes degenerate into an uninformed implicit mapping, forcing the model to blindly search an overwhelmingly vast hypothesis space. Ultimately, this lack of architectural regularization not only renders inference uninterpretable, but also provokes mutual interference between degradation removal and feature integration, frequently yielding physically implausible results that violate the intrinsic task dependency.

These limitations of both paradigms point to a common imperative: the network architecture should not be a heuristic assembly of generic operators, nor should it instantiate a biased decomposition of the joint task. Instead, it should faithfully reflect the task-related prior encoded in the optimization objective itself. A principled way to achieve this is to derive the architecture directly from an iterative algorithm that solves the joint objective. This approach, known as deep unfolding, provides an architectural scaffold that is both interpretable and task-prior-guided, guiding the restoration and fusion processes to evolve jointly toward a mutually consistent solution.

Guided by these motivations, we propose MODULE, a Mutual-prOmoting Deep UnfoLding framEwork that synergistically integrates restoration and fusion, as Fig. 1 (c) shows. First, we reformulate the restoration and fusion processes into a single joint optimization problem. This mathematical formulation reveals the intrinsic interdependence between two tasks, laying the foundation for algorithmic transparency. Second, the complex multi-task optimization is decomposed into a series of iterative, single-task subproblems, where the solution of each subproblem is guided by others. The iterative nature establishes a reciprocal information flow, steering the system toward a mutually consistent and physically plausible solution. Finally, we unfold the optimization process into a multi-stage deep network architecture. Constructed directly upon an optimization-based task-specific prior, MODULE not only inherits the interpretability of model-based methods but also leverages the representational capacity of networks for superior performance. Our contributions are as follows:

1. We reformulate degradation-robust image fusion as a joint optimization problem from a generative perspective. By explicitly modeling the dependency between restoration and fusion, our method effectively overcomes the limitations of cascaded suboptimal solutions and black-box opacity.

2. We decompose the complex multi-task optimization problem into a series of iterative, single-task subproblems, where the optimization of each is dynamically guided by the others. This iterative nature facilitates a bidirectional information flow, fostering mutual promotion that drives the system toward a mutually consistent and physically plausible solution.

3. We unfold the iterative optimization process into a multi-stage deep neural network. Each unfolded stage explicitly corresponds to a specific physical operation. It provides a transparent and interpretable architecture, seamlessly combining the advantages of model-based and data-driven approaches.

2 Related Work

Degradation-Robust Image Fusion. Achieving robust fusion against complex degradations remains challenging. Existing methods generally follow two paradigms. The cascaded paradigm integrates restoration and fusion; for example, URFusion [29] relies on high-quality priors to enhance degradation-invariant features, while DRMF [20] utilizes twin diffusion models for explicit pre-restoration. Conversely, the end-to-end paradigm implicitly learns robust mappings. Within this category, Text-IF [32] adopts language guidance to handle diverse degradations, and Deno-IF [28] separates clean components via low-rank decomposition. Yet, these methods fail to capture the inherent coupling between restoration and fusion, thereby preventing the two tasks from mutually regularizing and enhancing each other.

Deep Unfolding Networks. Deep unfolding networks (DUNs) bridge model-based optimization and deep learning via layer-wise iterative architectures. They are extensively explored in image restoration. ISTA-Net [36] unfolds shrinkage-thresholding iterations for compressive sensing. DGUNet [16] integrates gradient descent modules for image deblurring. VLU-Net [34] uses language prompts for multi-task recovery. IterRF [30] unfolds a unified registration–fusion optimization into alternating subproblems, enabling iterative bidirectional refinement of both tasks. Despite these advances, existing DUNs mainly focus on single-modal restoration or registration–fusion coordination, without explicitly addressing the coupled restoration and fusion of multiple modalities.

3 Method

3.1 Problem Formulation

In practical deployments, multi-modal sensors are susceptible to environmental variations, hardware imperfections, and operational inaccuracies. With i, v representing infrared and visible modalities, we can formulate the degraded source images based on the physical imaging process as:

$$y_m = D_m z_m + n_m,\ m \in \{i, v\}, \tag{2}$$

where y_m is the observed degraded source image. z_m is the corresponding high-quality image. D_m is the complex degradation operator (*e.g.*, blurring, low light). n_m represents noise.

Eq. (2) is a highly ill-posed inverse problem. Restoring z_m solely from the degraded observation y_m is inherently limited, hampering restoration performance. Fortunately, source images in the image

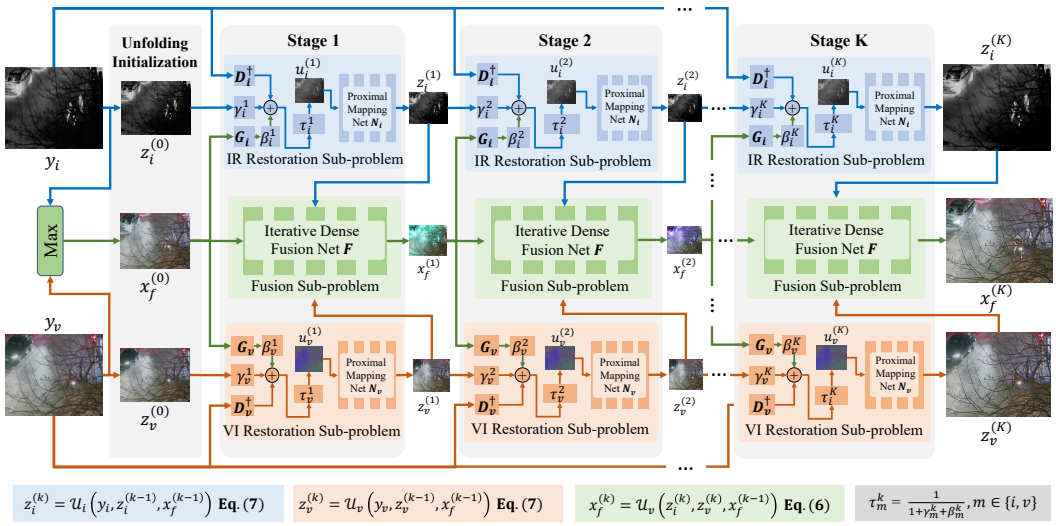


Figure 2: Overall network architecture of the proposed MODULE.

fusion task can provide complementary information about the same scene. The auxiliary modality offers a cross-modal prior to regularize the solution space, facilitating better restoration. However, challenges persist due to the significant discrepancy between different modalities. The irrelevant information in the auxiliary modality will interfere with the learning of correct mapping.

Thus, the core imperative is to maximize the guidance of cross-modal priors while suppressing cross-modal interference. It aligns with the goal of image fusion, *i.e.*, bridging modality gaps and aggregating informative features into a fused image x_f. The fusion process distills a purified, cross-modal prior, making x_f a comprehensive representation. Under this framework, two modalities are naturally modeled as distinct spectral projections of a unified prior. Thus, the high-quality source images can be generated by projecting the prior onto respective domains. From a generative perspective, the relationship is mathematically defined as:

$$z_m = G_m x_f,\ m \in \{i, v\}, \tag{3}$$

where G_m denotes a learnable modality-specific projection mapping, which projects the fused image back into source domains.

When linking the fusion prior to restored images, it raises a question: how to get x_f at first? Ideally, a high-quality x_f should be derived from z_m while generating z_m conversely relies on x_f and y_m. It reveals a circular dependency: estimating x_f requires z_m, and estimating z_m requires x_f. The interdependence cannot be resolved by sequentially performing fusion followed by restoration or vice versa. To this end, we unify fusion and restoration into a single, jointly optimized framework. It allows the two processes to mutually inform and constrain each other.

Mathematically, this problem can be formulated as a Maximum a Posteriori (MAP) estimation problem, *i.e.*, maximizing the posterior distribution $p\left(x_f, z_i, z_v | y_i, y_v\right)$. Under the common assumption, n_m can be reformulated as the quadratic data fidelity term through negative log-likelihood of MAP estimation. We transform the hard constraint of Eq. (3) into an optimizable soft penalty via the quadratic penalty method. Thus, x_f, z_m can be obtained by minimizing the following objective:

$$\min_{x_f, z_i, z_v} \sum\nolimits_{m \in \{i,v\}} \frac{1}{2}\|y_m - D_m z_m\|_F^2 + \lambda_m R_m(z_m) + \frac{\beta_m}{2}\|z_m - G_m x_f\|_F^2 + \lambda_f R_f(x_f), \tag{4}$$

where $\|y_m - D_m z_m\|_F^2$ and $\|z_m - G_m x_f\|_F^2$ are fidelity terms. $R_m(z_m)$ and $R_f(x_f)$ are regularization terms. R_m imposes prior constraints on the high-quality source images. R_f constrains the fused image. λ_m, β_m, and λ_f control the trade-off. This formulation narrows the solution space to a physically valid manifold, ensuring that recovered details are structurally consistent with the scene.

The original objective in Eq. (4) measures data fidelity in the observation space, where D_m may be severely ill-conditioned or non-invertible: the data term is nearly flat along the weakly observed directions and leaves the null-space components unconstrained, so the estimate there is driven by noise and by the regularizer alone. We therefore introduce the Moore–Penrose pseudo-inverse $D_m^\dagger$,

which is well defined for any bounded linear D_m even when its analytical form is unavailable, and measure the residual in the restoration space instead:

$$\min_{x_f,z_i,z_v} \sum\nolimits_{m\in\{i,v\}} \frac{1}{2}\|D_m^\dagger y_m - z_m\|_F^2 + \lambda_m R_m(z_m) + \frac{\beta_m}{2}\|z_m - G_m x_f\|_F^2 + \lambda_f R_f(x_f). \quad (5)$$

As analyzed in Appendix G, this reformulation yields an isotropic data term and explicitly penalizes the null-space component of z_m: among all z_m that explain the observation equally well, it selects the minimum-norm one and leaves the remaining null-space content to the prior. The ill-conditioned valley of the original data term thus becomes a well-conditioned one, and noise-driven variations in the unobservable directions are suppressed. It is a deliberate change of the objective rather than an equivalent rewriting of it.

In practice, D_m is unknown or lacks an explicit form; and even when available, a fixed linear inverse amplifies noise and cannot fill the unobservable directions. We therefore employ a prompt-driven neural network to learn a pseudo-inverse-inspired mapping that draws on image priors and cross-modal context to supply plausible structure where the observation carries none. Since it takes over the role of $D_m^\dagger y_m$ in Eq. (5), we retain this notation below without assuming any exact pseudo-inverse properties. Unlike a standalone restoration network, its output is consumed by the fusion branch and optimized under the joint objective, so it preserves what fusion needs rather than merely minimizing a restoration loss. We unfold the joint optimization into alternating subproblems and implement them as a deep network (Fig. 2).

3.2 Iterative Subproblem Decomposition

To solve this optimization problem, we use half-quadratic splitting (HQS) to decompose the complex multivariate (x_f, z_i, z_v) optimization problem into a series of iterative optimization sub-problems. The fusion sub-problem reconstructs the fused image $x_f^{(k)}$ from the current refined modal-specific images and previous fusion estimate, creating restoration-to-fusion promotion. With k denoting k-th optimization step, x_f-related (fusion) sub-problem can be formulated as:

$$x_f^{(k)} = \arg\min_{x_f} \frac{1}{2}\sum\nolimits_{m\in\{i,v\}} \beta_m \|z_m^{(k)} - G_m x_f\|_F^2 + \lambda_f R_f(x_f). \quad (6)$$

The restoration sub-problem estimates z_m from both observations and previously iterated fused result. z_m-related (restoration) sub-problem can be formulated as:

$$z_m^{(k)} = \arg\min_{z_m} \frac{1}{2}\|D_m^\dagger y_m - z_m\|_F^2 + \lambda_m R_m(z_m) + \frac{\beta_m}{2}\|z_m - G_m x_f^{(k-1)}\|_F^2, m \in \{i, v\}. \quad (7)$$

To resolve z_m-related sub-problem, we perform HQS to decouple the fidelity from the implicit image prior. By introducing auxiliary variable u to approximate z, u-subproblem is constructed as:

$$u_m^{(k)} = \arg\min_{u_m} \frac{1}{2}\|D_m^\dagger y_m - u_m\|_F^2 + \frac{\beta_m}{2}\|u_m - G_m x_f^{(k-1)}\|_F^2 + \frac{\gamma_m^{(k)}}{2}\|u_m - z_m^{(k-1)}\|_F^2, \quad (8)$$

where $\gamma_m^{(k)}$ is a penalty parameter. z-subproblem is constructed as:

$$z_m^{(k)} = \arg\min_{z_m} \lambda_m R_m(z_m) + \frac{\gamma_m^{(k)}}{2}\|z_m - u_m^{(k)}\|_F^2. \quad (9)$$

u_m enforces physical fidelity. z_m yields prior-enhanced reconstruction.

3.3 Optimization-Inspired Architecture Realization

To enable data-driven learning while maintaining interpretability, we instantiate each sub-problem as a dedicated network module. Mathematically, the u-subproblem admits a closed-form solution as:

$$u_m^{(k)} = \frac{D_m^\dagger y_m + \beta_m G_m x_f^{(k-1)} + \gamma_m^{(k)} z_m^{(k-1)}}{1 + \beta_m + \gamma_m^{(k)}}. \quad (10)$$

It acts as a fusion hub. Restoration is guided by fused $G_m x_f^{(k-1)}$ for fusion-to-restoration promotion.

z-subproblem is equivalent to a proximal mapping characterized by prior R_m. As a high-dimensional prior is difficult to model, we instantiate it with a network $\mathcal{N}_m$ to learn the manifold of high-quality images:

$$z_m^{(k)} = \mathcal{N}_m(u_m^{(k)}; \theta_k) \approx \text{prox}_{\frac{\lambda_m}{\gamma_m^{(k)}}}(u_m^{(k)}). \tag{11}$$

The fusion sub-problem aims to reconstruct the comprehensive scene x_f from the refined modality-specific images $z_i^{(k)}, z_v^{(k)}$. Instead of splitting x_f, we employ a fusion network $\mathcal{F}$ to learn high-dimensional feature aggregation directly. It avoids hand-crafted fusion rules and enables the model to distill "restoration-critical" features from $z_m^{(k)}$, realizing restoration-to-fusion promotion as:

$$x_f^{(k)} = \mathcal{F}(z_i^{(k)}, z_v^{(k)}, x_f^{(k-1)}). \tag{12}$$

Initialization. We initialize $z_m^{(0)} = y_m$. For the fused image, we adopt a pixel-wise maximum selection strategy, *i.e.*, $x_f^{(0)} = \max(y_i, y_v)$ to preserve salient thermal targets and structural textures across modalities, establishing a robust starting point for fast convergence during unfolding iterations.

Network Instantiation. We further instantiate the mathematical operators with specialized deep neural networks. The architecture is categorized into Operator Simulation Networks (instantiating $D_m^\dagger$ and G_m) and Solver Simulation Networks (instantiating $\mathcal{N}_m$ and $\mathcal{F}$). These networks are organized into K stages, sharing both the same architecture and learnable parameters.

Operator Simulation Networks. $D_m^\dagger$ is realized by a prompt-driven U-Net constructed by NAFBlocks [2]. Each block comprises a Prompt Generation Module (PGM) and a Prompt Interaction Module (PIM). PGM generates degradation-aware prompts from input features F. First, to capture global degradation context, we derive a content-adaptive weight vector w with global average pooling and gating mechanism as $w = \sigma(\text{Conv}_{gate}(\text{GAP}(F)))$, where σ is Sigmoid. Then, w modulates a learnable prompt bank $\mathbf{P}_{base}$, followed by a convolutional projection to obtain the specific prompt $P = \text{Conv}_{up}\left(w \odot \mathbf{P}_{base}\right)$. Then, PIM facilitates the interaction between features and prompts. By concatenating F and P, a NAFBlock refines features as $\hat{F} = \text{Conv}_{out}(\text{NAFBlock}([F, P]))$. G_m is instantiated as a network of stacked Restormer [33] blocks. With multi-head self-attention mechanism, G_m captures long-range dependencies, projecting the latent fused representation x_f back into modal-specific feature spaces while preserving spatial consistency. It should be emphasized that G_m is not required to reconstruct the full modality image; its role is to provide a modality-specific projection that transfers cross-modal structural priors to the restoration branch, and the hard relation $z_m = G_m(x_f)$ is relaxed into the soft consistency constraint L_p.

Solver Simulation Networks. Restoration solver $\mathcal{N}_m$, corresponding to z-subproblem, is constructed as a residual network embedded with squeeze-and-excitation blocks [5]. Through channel-wise attention, $\mathcal{N}_m$ dynamically recalibrates feature responses, emphasizing thermal targets and texture details. The fusion solver $\mathcal{F}$, corresponding to x_f-subproblem, adopts an asymmetric encoder-decoder. The encoder employs dense connections to maximize feature reuse. The decoder utilizes plain convolutions. For restoration-to-fusion promotion, we use a residual interaction from previous stage to learn incremental refinement of scene representation, ensuring stability and high-frequency preservation. β_m, γ_m are stage-wise learnable to adaptively balance data fidelity and prior regularization.

3.4 Loss Function

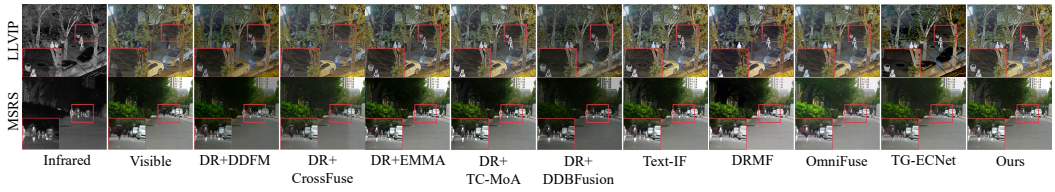
The overall loss function consists of two components: i) fusion loss $\mathcal{L}_{fuse}$ to regulate the fused image x_f, and ii) consistency loss $\mathcal{L}_{cons}$ to constrain the intermediate variables in iterative unfolding process, finally formulated as $\mathcal{L}_{total} = \mathcal{L}_{fuse} + \mathcal{L}_{cons}$.

Fusion Loss. To ensure that x_f preserves salient targets, fine textures, and natural color, we define $\mathcal{L}_{fuse}$ as a combination of intensity, gradient, and color constraints as:

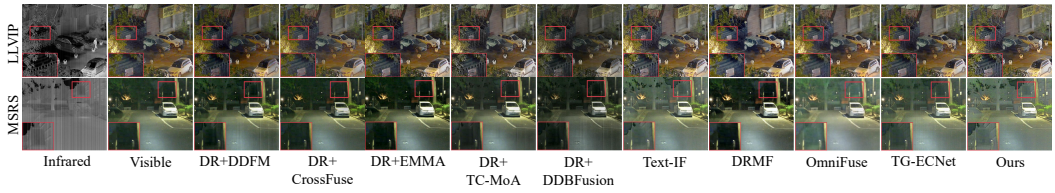
$$\mathcal{L}_{fuse} = \alpha_1\mathcal{L}_{int} + \alpha_2\mathcal{L}_{grad} + \alpha_3\mathcal{L}_{color}, \tag{13}$$

where $\alpha_1, \alpha_2, \alpha_3$ are hyperparameters to control trade-off.

The intensity loss $\mathcal{L}_{int}$ measures the intensity consistency between x_f and high-quality images I_v, I_i as $\mathcal{L}_{int} = \frac{1}{HW}\|x_f - \max(I_v, I_i)\|_1$. H, W are height and width. With the Sobel operator $\nabla(\cdot)$, the gradient loss $\mathcal{L}_{grad}$ constrains the alignment of edges as $\mathcal{L}_{grad} = \frac{1}{HW}\|\nabla x_f - \max(\nabla I_v, \nabla I_i)\|_1$.



(a) Qualitative Results on Infrared Gaussian Noise



(b) Qualitative Results on Infrared Stripe Noise

Figure 3: Qualitative results when the infrared image suffers from degradations.

The color loss $\mathcal{L}_{color}$ constrains color consistency as $\mathcal{L}_{color} = \frac{1}{HW}\|\mathcal{F}_{CbCr}(x_f) - \mathcal{F}_{CbCr}(I_{vis})\|_1$ where $\mathcal{F}_{\text{CbCr}}(\cdot)$ extracts the Cb and Cr chrominance channels.

Consistency Loss. This loss enforces the physical and mathematical alignment across the restoration and fusion tasks within all K stages:

$$\mathcal{L}_{cons} = \sum_{k=1}^{K} \sum_{m\in\{i,v\}} \lambda_z^{(k)}\mathcal{L}_z^{(k)} + \lambda_u^{(k)}\mathcal{L}_u^{(k)} + \lambda_p^{(k)}\mathcal{L}_p^{(k)}. \tag{14}$$

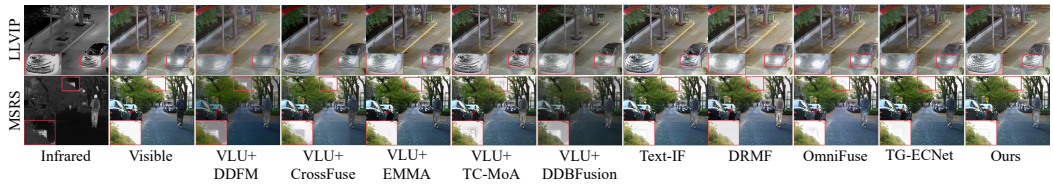
Restoration fidelity loss $\mathcal{L}_z^{(k)} = \|z_m^{(k)} - I_m\|_1$ supervises z_m to high-quality domain. With hyper-parameter η, restoration loss of final z_m is $\mathcal{L}_z^K = \mathcal{L}_z + \eta(1 - SSIM(z_m, I_m))$ to further preserve structure similarity in final iteration. Proximal alignment loss $\mathcal{L}_u^{(k)} = \|u_m^{(k)} - z_m^{(k)}\|_1$ ensures consistency between analytical solution u_m (derived from physics-based sub-problem) and z_m, maintaining integrity of HQS decomposition. Projection consistency loss $\mathcal{L}_p^{(k)} = \|G_m(x_f^{(k)}) - z_m^{(k)}\|_1$ enforces structural alignment between projection of x_f and restored image for fusion-restoration promotion.

4 Experiments and Results

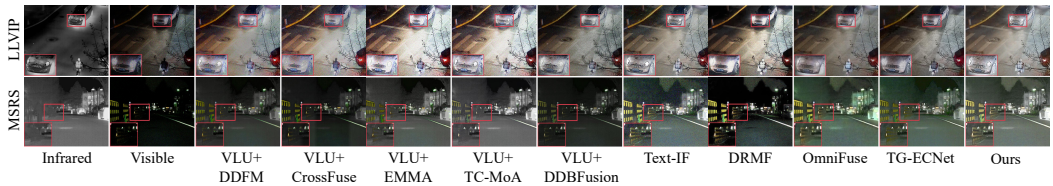
Datasets and Settings. We perform experiments on LLVIP [6] and MSRS [18] datasets. For each dataset, training and testing sets comprise 800 and 200 degraded–clean image pairs. Five types of degradations (infrared Gaussian noise, infrared stripe noise, visible Gaussian noise, visible low-light conditions, and visible defocus blur) are considered. Degradation parameters are randomly sampled per pair; details are in Appendix D. Experiments are conducted on an NVIDIA A800 GPU with PyTorch. We use the Adam optimizer, setting the initial learning rate to 0.001 with linear decay. Images are randomly cropped into 64×64. Epoch is 1500. Batch size is 24. K is 3. α_1, α_2, α_3 are 6, 8, 12. λ_u and λ_p are 1.5 and 2.0 and increase by 0.1 per iteration. λ_z is 1 for the first two iterations and 4 for final iteration. η is 1. Hyperparameters are empirically determined.

Multi-modal Image Fusion in Degraded Scenarios. Competitors include DDFM [39], CrossFuse [8], EMMA [40], TC-MoA [41], DDBFusion [37], Text-IF [32], DRMF [20], OmniFuse [35], and TG-ECNet [17]. For typical fusion methods, degraded infrared images are preprocessed by degradation-robust diffusion model in DRMF. Degraded visible images are preprocessed by an all-in-one restoration method VLU-Net [34]. Note that Text-IF, DRMF, OmniFuse and TG-ECNet remove degradations and fuse end-to-end.

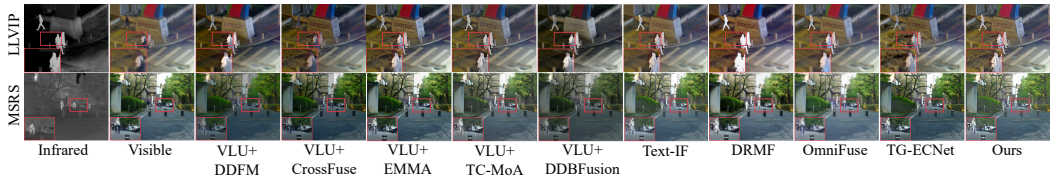
***i) Qualitative Comparison*.** Results when the infrared or visible image suffers from degradations are shown in Figs. 3-4, respectively. Our framework demonstrates integrated advantages in restoration and fusion via its coupled modeling and bidirectional guidance. For restoration, the fused image serves as the cross-modal prior that removes degradations, and the physically plausible modeling keeps all variables reciprocally regularized; in contrast, competing methods leave noticeable artifacts or residuals owing to their cascaded or black-box nature. For fusion, the high-fidelity restored results supply complementary information from both modalities, and the generative projection constraint of the unfolding network preserves salient thermal targets and fine textures.



(a) Qualitative Results on Visible Gaussian Noise



(b) Qualitative Results on Visible Low Light



(c) Qualitative Results on Visible Defocus Blur

Figure 4: Qualitative results when the visible image suffers from degradations.

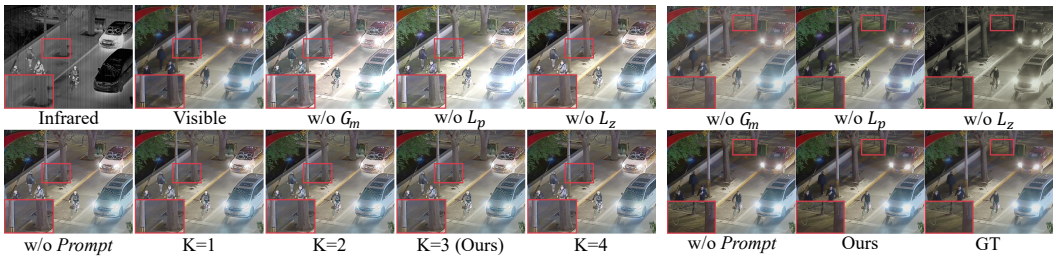
Table 1: Quantitative results on degraded infrared and visible image fusion (best, second-best).

Datasets	**LLVIP Dataset**								**MSRS Dataset**							
Degradations	*Infrared Degradations*				*Visible Degradations*				*Infrared Degradations*				*Visible Degradations*			
Methods	SCD↑	VIF↑	Q^{abf}↑	SSIM↑	SCD↑	VIF↑	Q^{abf}↑	SSIM↑	SCD↑	VIF↑	Q^{abf}↑	SSIM↑	SCD↑	VIF↑	Q^{abf}↑	SSIM↑
RM+DDFM	1.293	0.400	0.301	0.650	1.575	0.445	0.338	0.683	1.282	0.480	0.289	0.527	1.354	0.654	0.300	0.667
RM+CrossFuse	0.754	0.355	0.307	0.651	0.831	0.420	0.286	0.530	0.963	0.487	0.353	0.669	0.959	0.518	0.276	0.528
RM+TC-MoA	1.181	0.414	0.374	0.699	1.287	0.478	0.399	0.663	1.293	0.542	0.374	0.683	1.257	0.765	0.392	0.760
RM+EMMA	1.283	0.361	0.333	0.663	1.374	0.447	0.330	0.602	1.353	0.516	0.362	0.662	1.315	0.613	0.337	0.628
RM+DDBFusion	1.255	0.359	0.246	0.607	1.531	0.402	0.280	0.656	1.229	0.402	0.241	0.473	1.299	0.598	0.271	0.644
Text-IF	1.322	0.464	0.453	0.739	1.266	0.466	0.422	0.648	1.378	0.480	0.345	0.443	1.291	0.552	0.341	0.501
DRMF	1.078	0.395	0.322	0.644	0.898	0.410	0.275	0.421	1.214	0.505	0.322	0.598	1.153	0.445	0.257	0.382
OmniFuse	1.220	0.366	0.255	0.602	1.166	0.420	0.244	0.539	1.167	0.519	0.280	0.498	1.146	0.568	0.291	0.508
TG-ECNet	1.083	0.356	0.313	0.669	1.132	0.404	0.292	0.546	0.361	0.497	0.352	0.511	1.205	0.541	0.309	0.567
Ours	1.480	0.455	0.462	0.773	1.448	0.575	0.498	0.831	1.389	0.621	0.395	0.788	1.345	0.769	0.451	0.782

ii) Quantitative Comparison. Evaluation metrics include SCD [1], VIF [4], Q_{abf} [13] and SSIM [24]. Preprocessing methods are referred to as Restoration Methods (RM) for simplicity. In Tab. 1, existing methods are constrained by cascaded pipelines or black-box designs and fail to disentangle valid scenes. In contrast, our method achieves best or second-best results on most metrics across datasets and degradations. Per-degradation results are given in Appendix A.

Ablation Study. We ablate the proposed components, the unfolding depth and the mutual promotion between the two branches. Removing G_m, L_p, L_z or the prompt block in $D^\dagger$ one at a time causes overexposure, color shifts, texture distortion and slight texture loss, and degrades most metrics in Tab. 2; increasing the unfolding depth improves performance up to three stages and slightly degrades it at four while the training losses in Tab. 12 (Appendix C) keep decreasing, so we adopt $K = 3$. For restoration to fusion promotion, the fused output in Fig. 5(a) deteriorates whenever the restoration branch is ablated, and the corresponding drops in Tab. 2 confirm that better restoration benefits fusion. For fusion to restoration promotion, disabling the feedback makes the intermediate visible restoration lose the color and structure of the low-light scene, whereas the full model recovers them from the intact infrared structure in Fig. 5(b); the convergence of the iterative variables in Tab. 5 and the quantitative evidence in Appendix C show that fusion in turn guides restoration.

Generalization to Unseen and Compound Degradations. We test generalization under two conditions absent from training: an unseen dataset with both modalities degraded, and a compound degradation combining two corruptions of the visible image. For the unseen setting, we evaluate on 100 M3FD pairs never involved in training and add Gaussian noise to both modalities, stripe noise to



(a) Ablation qualitative results on the fused image. **(b) Ablation qualitative results on z_v.**

Figure 5: Qualitative results of ablation study.

Table 2: Ablation studies on the LLVIP dataset.

Variant	Infrared				Visible			
	SCD ↑	VIF ↑	Q^{abf} ↑	SSIM ↑	SCD ↑	VIF ↑	Q^{abf} ↑	SSIM ↑
w/o G_m	1.296	0.374	0.370	0.642	1.299	0.589	0.446	0.751
w/o L_p	1.284	0.384	0.368	0.631	1.319	0.596	0.440	0.706
w/o L_z	1.546	0.457	0.448	0.745	1.490	0.529	0.443	0.782
w/o Prompt	1.480	0.434	0.440	0.772	1.477	0.590	0.498	0.874
$K = 1$	1.516	0.466	0.462	0.804	1.489	0.581	0.490	0.857
$K = 2$	1.518	0.467	0.460	0.801	1.503	0.593	0.499	0.872
$K = 3$ **(Ours)**	1.545	0.471	0.466	0.815	1.516	0.591	0.498	0.873
$K = 4$	1.539	0.472	0.465	0.809	1.500	0.585	0.494	0.869

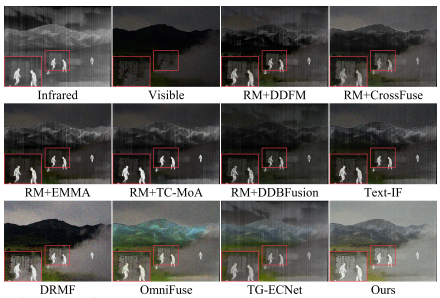


Figure 6: Results on the unseen M3FD dataset under cross-modal degradation.

Table 3: Evaluation on unseen and compound degradations (best, second-best).

(a) Unseen M3FD, both modalities degraded beyond the training range.

Method	SCD ↑	VIF ↑	Q^{abf} ↑	PSNR ↑	SSIM ↑
DDFM	1.345	0.307	0.177	12.487	0.228
CrossFuse	0.682	0.304	0.196	13.250	0.205
TC-MoA	1.034	0.328	0.222	13.686	0.236
EMMA	1.093	0.323	0.218	13.412	0.226
DDBFusion	1.286	0.264	0.156	12.197	0.202
Text-IF	1.002	0.193	0.177	12.657	0.127
DRMF	0.812	0.263	0.188	12.905	0.174
OmniFuse	1.216	0.384	0.212	14.291	0.280
TG-ECNet	1.255	0.318	0.208	14.968	0.235
Ours	1.512	0.532	0.348	14.345	0.501

(b) Compound degradation on LLVIP (low light and defocus).

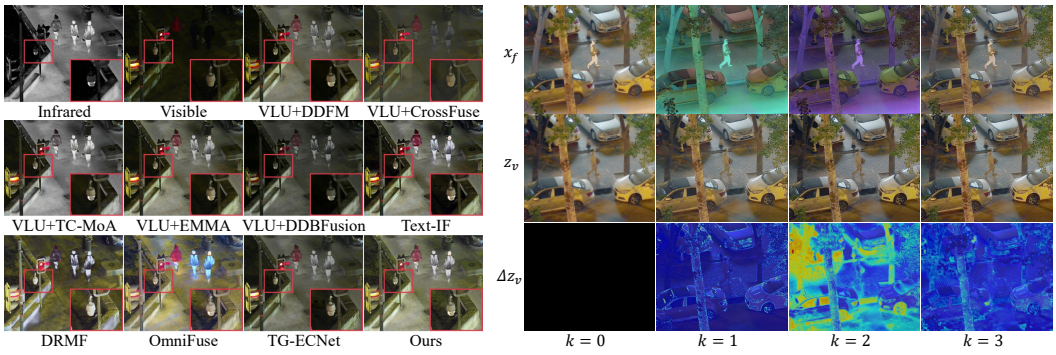
Method	SCD ↑	VIF ↑	Q^{abf} ↑	PSNR ↑	SSIM ↑
DDFM	1.254	0.376	0.168	14.013	0.372
CrossFuse	0.845	0.291	0.083	14.199	0.284
TC-MoA	1.185	0.484	0.195	14.013	0.422
EMMA	1.178	0.387	0.179	14.267	0.354
DDBFusion	1.056	0.329	0.138	12.944	0.325
Text-IF	1.109	0.222	0.181	15.416	0.198
DRMF	0.838	0.229	0.114	13.974	0.189
OmniFuse	1.148	0.317	0.131	16.364	0.359
TG-ECNet	1.187	0.367	0.175	14.322	0.347
Ours	1.333	0.398	0.196	15.918	0.488

the infrared image, and low light with shot noise together with defocus blur to the visible image, with all strengths beyond the training ranges. Fig. 6 and Tab. 3(a) show that most baselines either leave the noise and stripe patterns visible or suppress them at the price of washed-out thermal targets and lost visible textures, whereas MODULE keeps the thermal signatures sharp and the visible textures intact, ranking first on four of the five metrics and second on the remaining one; the only metric led by a baseline is PSNR, reflecting our emphasis on structural and perceptual fidelity over pixel-wise reconstruction under heavy simultaneous degradation. Fig. 8 in Appendix B and Tab. 3(b) show the same under compound degradation: MODULE preserves structural fidelity and gradient information even when luminance and sharpness cues are removed together.

Extensive Experiments. We conduct extensive experiments including real-world validation, interpretability and convergence analysis, high-level validation and complexity analysis.

***i) Real-world Validation*.** We select degraded images (scarce in datasets) from LLVIP and MSRS datasets, 45 in total and none of which is involved in training. Fig. 7 (a) shows that our framework retains vital texture information while successfully enhancing the illumination of the visible image. Tab. 4 demonstrates that our framework outperforms all competitors in non-reference metrics.

ii) Interpretability and Convergence Analysis. Fig. 7 (b) visualizes fused image x_f, restored visible image z_v and iterative difference maps Δz_v under visible defocus blur. The variables evolve in a coarse-to-fine manner, progressively eliminating ghosting and recovering details. It confirms our



(a) Qualitative results on real-world scenarios. **(b) Intermediate variable visualization on visible defocus blur.**

Figure 7: Qualitative results of real-world validation and intermediate variable visualization.

Table 4: Experiment on real-world images and convergence analysis of intermediate variables.

Methods	DDFM	CrossFuse	TC-MoA	EMMA	DDBFusion	Text-IF	DRMF	OmniFuse	TG-ECNet	**Ours**
NIQE ↓	4.360	4.253	4.200	4.668	4.040	4.198	5.115	6.057	5.069	3.667
BRISQUE ↓	31.590	27.927	26.616	25.148	23.288	28.663	41.255	47.784	24.457	22.356
MUSIQ ↑	43.046	44.482	45.368	44.385	43.332	46.212	49.835	42.269	47.670	50.122

Table 5: Convergence analysis under representative degradations (lowest, second-lowest).

Iteration k	**Infrared: Gaussian Noise**				**Visible: Low Light**			
	$Z_i - I_i$	$G_i - I_i$	$U_i - Z_i$	$L(x_f)$	$Z_v - I_v$	$G_v - I_v$	$U_v - Z_v$	$L(x_f)$
$k=1$	0.050	0.194	0.009	0.121	0.064	0.229	0.054	0.189
$k=2$	0.013	0.137	0.009	0.115	0.024	0.069	0.046	0.161
$k=3$	0.011	0.136	0.009	0.115	0.022	0.048	0.045	0.157

Table 6: Quantitative comparison of object detection performance and computational complexity.

Detection Metrics	DDFM	CrossFuse	TC-MoA	EMMA	DDBFusion	Text-IF	DRMF	OmniFuse	TG-ECNet	**Ours**
Precision	0.943	0.877	0.906	0.922	0.929	0.898	0.917	0.897	0.908	0.934
Recall	0.731	0.699	0.721	0.739	0.714	0.765	0.678	0.728	0.743	**0.767**
AP@0.5	0.821	0.806	0.808	0.822	0.816	0.843	0.798	0.798	0.829	**0.867**
AP@[.5:.95]	0.544	0.478	0.521	0.528	0.546	0.556	0.513	0.512	0.561	**0.584**
Complexity and Efficiency	DDFM	CrossFuse	EMMA	TC-MoA	DDBFusion	Text-IF	DRMF	OmniFuse	TG-ECNet	**Ours**
Parameter Numbers (M) ↓	552.16	23.15	5.06	115.40	9.10	215.12	170.90	98.18	160.86	16.09
Time (s) ↓	63.38	3.55	2.89	1.43	2.83	1.22	13.22	5.28	1.93	0.63
FLOPs (G) ↓	4079	23	42	115	1664	215	171	5470	1998	317

framework is transparent and task-guided. Tab. 5 shows all intermediate variables converge to ideal targets during inference, indicating that our framework simulates the optimization algorithm.

iii) High-level Validation and Complexity Analysis. We evaluate object detection performance in low-light scenarios on LLVIP with YOLOv11 [7] as detector. Tab. 6 reveals that MODULE achieves SOTA performance across all metrics as it can balance detail preservation and degradation suppression, achieving robustness in complex scenarios. In terms of computational cost, MODULE is lightweight in both parameters and FLOPs, so its multi stages do not sacrifice efficiency.

5 Conclusion

We propose a deep unfolding framework that reformulates degradation-robust fusion as a task-prior-guided joint optimization, eliminating cascaded suboptimal solution and black-box opacity. By decomposing the optimization into iterative subproblems, it establishes bidirectional mutual promotion within a transparent architecture, bridging model-based prior with data-driven power.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62502085 and U21B2027, the Natural Science Foundation of Jiangsu Province of China under Grant BK20241280, and the Start-up Research Fund of Southeast University under Grant RF1028624061 and RF1028623006. The authors declare no competing financial interests.

References

[1] V. Aslantas and 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.

[2] L. Chen, X. Chu, X. Zhang, and J. Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022.

[3] C. Cheng, T. Xu, Z. Feng, X. Wu, Z. Tang, H. Li, Z. Zhang, S. Atito, M. Awais, and J. Kittler. One model for all: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28102–28112, 2025.

[4] Y. Han, Y. Cai, Y. Cao, and X. Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013.

[5] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[6] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.

[7] R. Khanam and M. Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

[8] H. Li and X.-J. Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024.

[9] H. Li, Z. Yang, Y. Zhang, W. Jia, Z. Yu, and Y. Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3673–3690, 2025.

[10] J. Liu, G. Wu, Z. Liu, D. Wang, Z. Jiang, L. Ma, W. Zhong, X. Fan, and R. Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2024.

[11] J. Liu, B. Zhang, Q. Mei, X. Li, Y. Zou, Z. Jiang, L. Ma, R. Liu, and X. Fan. 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.

[12] H. Ma, H. Li, C. Cheng, G. Wang, X. Song, and X.-J. Wu. S4fusion: Saliency-aware selective state space model for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 34:4161–4175, 2025.

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

[14] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

[15] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2020.

[16] C. Mou, Q. Wang, and J. Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17399–17410, 2022.

[17] Y. Sun, X. Li, P. Zhu, Q. Hu, D. Ren, H. Xu, and X. Zhu. Task-gated multi-expert collaboration network for degraded multi-modal image fusion. In *International Conference on Machine Learning*, pages 57571–57586. PMLR, 2025.

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

[19] L. Tang, H. Zhang, H. Xu, and J. Ma. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, 99:101870, 2023.

[20] L. Tang, Y. Deng, X. Yi, Q. Yan, Y. Yuan, and J. Ma. Drmf: Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8546–8555, 2024.

[21] L. Tang, C. Li, and J. Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):591–608, 2025.

[22] W. Tang, F. He, Y. Liu, Y. Duan, and T. Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7): 3159–3172, 2023.

[23] Z. Tang, T. Xu, X.-J. Wu, and J. Kittler. Multi-level fusion for robust rgbt tracking via enhanced thermal representation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(10):1–24, 2024.

[24] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9 (3):81–84, 2002.

[25] X. Wu, Z.-H. Cao, T.-Z. Huang, L.-J. Deng, J. Chanussot, and G. Vivone. Fully-connected transformer for multi-source image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2071–2088, 2025.

[26] X. Xie, Y. Cui, T. Tan, X. Zheng, and Z. Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1):37, 2024.

[27] H. Xu, Y. Huang, L. Tang, J. Ma, and G. Liu. Diff-mef: Cross-modal diffusion framework with text prompts and semantic perception for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 35:3186–3201, 2025.

[28] H. Xu, Y. Li, Y. Deng, J. Ma, and G. Liu. Deno-if: Unsupervised noisy visible and infrared image fusion method. In *Annual Conference on Neural Information Processing Systems*, volume 38, pages 141649–141671, 2025.

[29] H. Xu, X. Yi, C. Lu, G. Liu, and J. Ma. Urfusion: Unsupervised unified degradation-robust image fusion network. *IEEE Transactions on Image Processing*, 34:5803–5818, 2025.

[30] H. Xu, J. Chen, X. Yi, P. Li, and J. Ma. Iterrf: Iterative and interpretable mutual enhancement for joint multi-modal image registration and fusion. *International Journal of Computer Vision*, 134(10):442, 2026.

[31] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.

[32] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024.

[33] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[34] H. Zeng, X. Wang, Y. Chen, J. Su, and J. Liu. Vision-language gradient descent-driven all-in-one deep unfolding networks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7524–7533, 2025.

[35] H. Zhang, L. Cao, X. Zuo, Z. Shao, and J. Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025.

[36] J. Zhang and B. Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1828–1837, 2018.

[37] Z. Zhang, H. Li, T. Xu, X.-J. Wu, and J. Kittler. Ddbfusion: An unified image decomposition and fusion framework based on dual decomposition and bézier curves. *Information Fusion*, 114: 102655, 2025.

[38] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool. 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.

[39] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023.

[40] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, and L. Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25912–25921, 2024.

[41] P. Zhu, Y. Sun, B. Cao, and Q. Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7099–7108, 2024.

Technical Appendices and Supplementary Material

A Quantitative Results of Each Degradation Type

Results on Infrared Image Restoration and Fusion. As reported in Table 7, our method consistently outperforms existing state-of-the-art benchmarks in handling Gaussian and Stripe noise within the infrared modality. Specifically, our approach achieves the highest SCD and VIF scores in most cases, indicating superior information transfer and visual fidelity. Notably, for structured degradations like stripe noise, which often perplexes pure data-driven models, our unfolding architecture demonstrates remarkable robustness by effectively identifying and suppressing directional artifacts while preserving the thermal radiation characteristics of targets.

Table 7: Quantitative results of infrared image restoration and fusion under Gaussian and Stripe noise (best, second-best).

Method	Gaussian Noise								Stripe Noise							
	LLVIP				MSRS				LLVIP				MSRS			
	SCD	VIF	Q_{abf}	SSIM	SCD	VIF	Q_{abf}	SSIM	SCD	VIF	Q_{abf}	SSIM	SCD	VIF	Q_{abf}	SSIM
DDFM	1.246	0.408	0.302	0.663	1.293	0.493	0.301	0.671	1.341	0.392	0.300	0.638	1.272	0.466	0.278	0.383
CrossFuse	0.657	0.355	0.303	0.648	0.937	0.505	0.377	0.897	0.851	0.355	0.310	0.654	0.988	0.470	0.329	0.440
TC-MoA	1.110	0.417	0.372	0.707	1.322	0.560	0.394	0.899	1.252	0.410	0.375	0.691	1.263	0.523	0.354	0.466
EMMA	1.244	0.364	0.332	0.668	1.334	0.532	0.383	0.885	1.321	0.359	0.334	0.658	1.372	0.500	0.342	0.440
DDBFusion	1.214	0.368	0.247	0.620	1.244	0.411	0.248	0.600	1.296	0.351	0.245	0.595	1.214	0.393	0.235	0.347
Text-IF	1.256	0.460	0.441	0.723	1.310	0.391	0.298	0.365	1.388	0.469	0.465	0.755	1.446	0.568	0.392	0.520
DRMF	1.035	0.397	0.320	0.646	1.198	0.522	0.344	0.800	1.120	0.393	0.325	0.643	1.230	0.488	0.299	0.397
OmniFuse	1.157	0.359	0.246	0.593	1.178	0.526	0.285	0.626	1.283	0.374	0.264	0.611	1.156	0.512	0.276	0.371
TG-ECNet	0.843	0.362	0.314	0.675	0.349	0.519	0.375	0.568	1.323	0.351	0.312	0.663	0.373	0.475	0.328	0.455
Ours	1.473	0.447	0.433	0.771	1.380	0.586	0.394	0.853	1.488	0.462	0.490	0.775	1.398	0.655	0.397	0.723

Results on Visible Image Restoration and Fusion. Table 8 presents the quantitative comparison under more challenging visible degradations, including low light and defocus blur. Our method remains competitive across all metrics, with the largest gains on structural similarity and gradient fidelity, reflecting its exceptional ability to reconstruct structural details and maintain edge information. The significant performance gain under low-light and blurred conditions stems from our bidirectional interaction mechanism, which adaptively leverages the sharp structural guidance from the infrared branch to compensate for the lost textures in the degraded visible branch.

Table 8: Results of visible image restoration and fusion under Low Light and Defocus Blur.

Method	Low Light								Defocus Blur							
	LLVIP				MSRS				LLVIP				MSRS			
	SCD	VIF	Q_{abf}	SSIM	SCD	VIF	Q_{abf}	SSIM	SCD	VIF	Q_{abf}	SSIM	SCD	VIF	Q_{abf}	SSIM
DDFM	1.620	0.365	0.343	0.515	1.158	0.617	0.285	0.539	1.530	0.447	0.272	0.707	1.413	0.685	0.243	0.691
CrossFuse	0.825	0.323	0.305	0.414	0.837	0.469	0.241	0.367	0.856	0.474	0.169	0.523	1.068	0.590	0.241	0.557
TC-MoA	1.241	0.369	0.335	0.436	1.070	0.812	0.384	0.728	1.224	0.500	0.342	0.734	1.306	0.783	0.333	0.755
EMMA	1.302	0.364	0.301	0.443	1.039	0.612	0.317	0.515	1.339	0.494	0.250	0.627	1.414	0.660	0.296	0.636
DDBFusion	1.537	0.320	0.296	0.474	1.013	0.523	0.243	0.481	1.532	0.415	0.221	0.686	1.431	0.666	0.243	0.709
Text-IF	1.372	0.404	0.417	0.510	1.323	0.427	0.342	0.330	1.201	0.496	0.358	0.720	1.318	0.841	0.344	0.757
DRMF	0.988	0.357	0.302	0.380	1.167	0.393	0.273	0.254	0.867	0.443	0.144	0.381	1.129	0.545	0.207	0.434
OmniFuse	1.216	0.467	0.293	0.603	1.196	0.616	0.352	0.520	1.086	0.415	0.145	0.450	1.119	0.589	0.202	0.436
TG-ECNet	1.162	0.431	0.316	0.553	1.070	0.482	0.292	0.416	1.093	0.380	0.210	0.489	1.155	0.593	0.258	0.572
Ours	1.416	0.576	0.490	0.793	1.407	0.818	0.480	0.768	1.385	0.611	0.488	0.829	1.374	0.829	0.466	0.835

Table 9 presents the quantitative comparison under visible Gaussian noise. Our visible restoration branch can handle blur, which tends to recover high-frequency details. In contrast, Gaussian denoising requires aggressive smoothing that opposes the objective of deblurring's structural preservation. Despite the gradient conflict of deblurring and denoising tasks, our framework maintains competitive performance on both datasets.

In summary, the comprehensive quantitative results across diverse datasets and degradation scenarios validate the effectiveness of our deep unfolding framework.

Table 9: Quantitative results on Visible Gaussian Noise degradation (best, second-best).

Gaussian Noise	LLVIP				MSRS			
	SCD	VIF	Qabf	SSIM	SCD	VIF	Qabf	SSIM
VLU+DDFM	1.574	0.522	0.398	0.826	1.490	0.660	0.371	0.770
VLU+CrossFuse	0.812	0.462	0.383	0.654	0.972	0.494	0.345	0.661
VLU+TC-MoA	1.395	0.567	0.522	0.820	1.395	0.699	0.460	0.798
VLU+EMMA	1.480	0.484	0.439	0.736	1.491	0.568	0.399	0.732
VLU+DDBFusion	1.523	0.470	0.324	0.806	1.451	0.604	0.326	0.743
Text-IF	1.226	0.499	0.490	0.716	1.231	0.387	0.336	0.417
DRMF	0.840	0.430	0.378	0.501	1.162	0.397	0.291	0.459
OmniFuse	1.197	0.378	0.295	0.565	1.122	0.498	0.318	0.569
TG-ECNet	1.141	0.401	0.350	0.597	1.391	0.549	0.376	0.712
Ours	1.544	0.538	0.519	0.869	1.254	0.661	0.406	0.744

B Compound Degradation on LLVIP

This setting uses 100 image pairs from LLVIP, where low-light corruption and defocus blur are applied simultaneously to the visible image while the infrared image stays intact. This combination never appears during training, so the test probes whether the framework handles two coupled corruptions rather than a single degradation family; results are given in Fig. 8 and Tab. 3(b) of the main text.

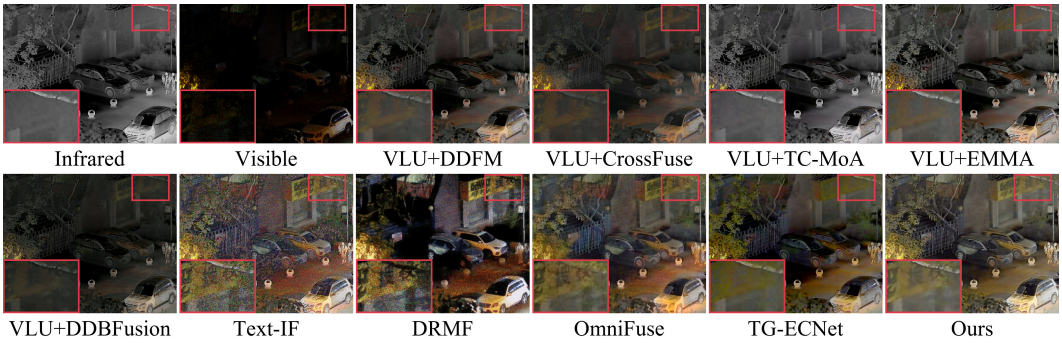


Figure 8: Results under compound degradation on LLVIP, where low light and defocus blur are applied to the visible image simultaneously. The insets enlarge the red-boxed regions.

Fig. 8 shows that the baselines over-brighten the scene while leaving residual blur, or sharpen edges at the cost of amplified noise and color distortion; MODULE instead recovers the lost sharpness from the intact infrared structure without amplifying the compounded artifacts. Tab. 3(b) shows the same trend: MODULE attains the best structural similarity, information transfer and gradient fidelity and remains among the top performers in the remaining metrics, while the baselines lose more accuracy as the corruption departs from the training distribution. The advantage in SSIM indicates that removing luminance and sharpness cues together does not disrupt the cross-modal structural prior exploited by the fusion branch.

C Additional Ablation Studies

This section reports three additional ablations that the main text omits for space: the initialization strategy of the iterative process (Tab. 10), the fidelity constraint L_u (Tab. 11), and the training behaviour with respect to the unfolding depth (Tab. 12).

i) Initialization Strategy. Tab. 10 shows that replacing the pixel-wise maximum selection by averaging or a learnable convolution does not improve the final fusion, and the maximum rule is consistently the best. This is expected: only a structural initialization is required for the first stage, and the element-wise maximum preserves the salient response of either modality without introducing additional parameters.

ii) Fidelity Constraint L_u. Tab. 11 shows that the fidelity constraint mainly benefits the visible branch, which is the more severely degraded one: removing it degrades both the fused output and the

restored visible image, and the perceptual distance grows markedly. On the infrared branch the effect is marginal and the difference in one metric falls within the rounding level, which is consistent with the observation that an already mild degradation leaves little room for the constraint to act.

Table 10: Effect of the initialization strategy of the iterative process (best, second-best).

Setting	Initialization	SCD ↑	VIF ↑	Q^{abf} ↑	SSIM ↑
IR Gaussian	Mean	1.449	0.400	0.407	0.731
	Conv	1.397	0.411	0.422	0.765
	Max (Ours)	1.473	0.447	0.433	0.771
VIS low light	Mean	1.405	0.572	0.475	0.789
	Conv	1.409	0.563	0.462	0.780
	Max (Ours)	1.416	0.576	0.490	0.793

Table 11: Effect of the fidelity constraint L_u on fusion and restoration (best).

Setting	Variant	SCD ↑	VIF ↑	Q^{abf} ↑	PSNR ↑	SSIM ↑	LPIPS ↓
IR Gaussian	w/o L_u	1.476	0.440	0.432	36.324	0.943	0.088
	Ours	1.473	0.447	0.433	36.422	0.944	0.087
VIS low light	w/o L_u	1.406	0.567	0.470	27.123	0.883	0.160
	Ours	1.416	0.576	0.490	27.377	0.891	0.138

iii) Unfolding Depth and Training Losses. Deeper unfolding keeps reducing the training losses in Tab. 12 (L_{fuse} and the projection, fidelity and restoration terms of both branches), yet the test performance in Tab. 2 peaks at three stages and slightly declines afterwards. This gap between training and test behaviour indicates that the additional stage starts to fit the synthetic training degradations rather than the underlying degradation model, which is why we adopt three stages as a compromise between representation capacity and generalization.

Table 12: Training losses of the final stage under different unfolding depths (lowest).

K	L_{fuse}	L_{pi}	L_{pv}	L_{ui}	L_{uv}	L_{zi}	L_{zv}
3	2.626	0.011	0.009	0.005	0.025	0.014	0.016
4	2.586	0.007	0.006	0.033	0.006	0.016	0.014

Table 13: Restoration quality under feedback removal.

Variant	**Visible**			**Infrared**		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o G_m	15.674	0.817	0.196	36.388	0.943	0.088
w/o L_p	13.545	0.790	0.233	36.401	0.943	0.082
Ours	27.377	0.891	0.138	36.422	0.944	0.087

iv) Restoration under Feedback Removal. Tab. 13 reports the restoration quality of the intermediate estimates when the fusion-to-restoration feedback is disabled. Disabling the feedback markedly degrades the visible restoration, and removing only the projection consistency loss degrades it further, since an unconstrained projection writes erroneous structure back. The infrared branch changes only marginally across the variants, which is consistent with its much milder degradation.

D Detailed Degradation Generation

This section describes how the synthetic degradations used throughout the paper are generated, including the formulation of each degradation type and the range of its parameters. All degraded–clean pairs are synthesized from clean LLVIP and MSRS images, and each pair is degraded exactly once, with every parameter independently sampled from a fixed distribution.

Gaussian noise (both modalities). $I_{noisy} = I + \mathcal{N}(0, \sigma^2)$ with $\sigma \sim U(5, 30)$. Noise is added to each RGB channel independently for visible image and to the single channel for infrared image.

Stripe noise (infrared). $I_{noisy}(x, y) = I(x, y) + S(y)$ with $S(y) \sim \mathcal{N}(0, \beta^2)$ and $\beta \sim U(3, 8)$; a single one-dimensional noise vector is shared by all rows, which reproduces the column-wise pattern of readout electronics.

Defocus blur (visible). $I_{blur} = I * K$, where K is a normalized circular point-spread function of size $k \times k$ with radius $r = \lfloor k/2 \rfloor$ and $k \in \{5, 7, 9, 11\}$.

Low light with shot noise (visible). The illumination is first attenuated by $\tilde{I} = (I/255)^{\gamma}$ with $\gamma \sim U(2.0, 3.5)$, and sensor noise is then injected following a Poisson model $I_{noisy} = (255/P) \cdot \mathcal{P}(\tilde{I} \cdot P)$ with $P \sim U(15, 35)$; a smaller P corresponds to a lower signal-to-noise ratio.

Training and testing pairs are drawn independently from these distributions. The compound setting in Appendix B combines two visible corruptions that never co-occur during training.

E Medical Evaluation

We select 100 MRI-PET image pairs from the Harvard Medical dataset[2] for training and 50 pairs for testing. For competitors, we select methods capable of performing medical image fusion. As Fig. 9 shows, our method exhibits a superior capability in structural and texture restoration, preserving sharp anatomical edges and fine-grained diagnostic details that are often blurred by other SOTA methods.

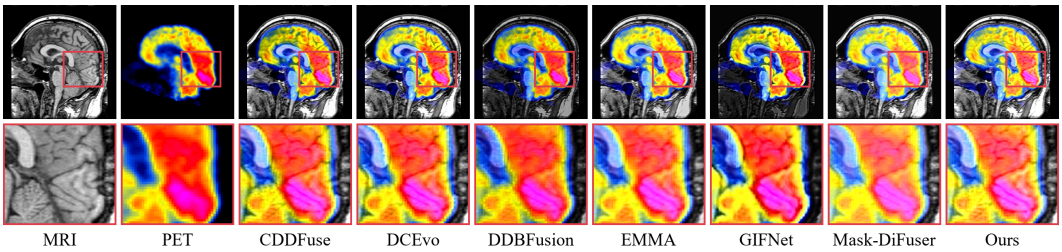


Figure 9: Qualitative results in MRI-PET image fusion.

As Tab. 14 shows, our method achieves advantages across all evaluated metrics. The results underscore the benefit of our generative modeling approach. By bypassing the limitations of hand-crafted modal-specific priors, our framework robustly handles the distinct imaging physics of medical modalities, providing a unified and high-performance solution for multi-modal image fusion.

Table 14: Quantitative comparison on medical image fusion (best, second-best).

Metrics	CDDFuse [38]	EMMA [40]	GIFNet [3]	DCEvo [11]	DDBFusion [37]	Mask-Difuser [21]	**Ours**
SCD ↑	1.529	1.471	1.296	1.478	1.271	1.619	1.616
VIF ↑	0.776	0.565	0.416	0.787	0.561	0.759	0.793
Q_{abf} ↑	0.751	0.610	0.339	0.742	0.508	0.694	0.753
SSIM ↑	0.632	0.370	0.225	0.373	0.503	0.373	0.726

F Coefficient Analysis

As the core contribution is the restoration-fusion mutual-promoting mechanism, we analyze coefficients of losses critical to this mechanism.

For fusion loss, the gradient loss preserves perceptually important structures and high-frequency textures, making α_2 highly influential for fusion performance. Moreover, via G_m, these structures are back-projected to guide restoration, making α_2 also critical for the joint system. Other coefficients are fixed to keep the output usable. For consistency loss, λ_p of L_p controls our novel inter-task mutual promotion mechanism. Analyzing it reveals the effect of this mechanism on performance and helps find a balanced regime. Other consistency losses are tied to the physical meanings of intermediate variables and thus fixed to avoid confounding.

As Fig. 10 shows, α_2 balances sharpness and fidelity: $\alpha_2 = 0$ loses high-frequency details, while $\alpha_2 = 50$ over-amplifies gradients, causing color distortion. The coupling term λ_p: $\lambda_p = 0$ decouples tasks, leading to color shift; $\lambda_p = 10$ over-restricts flexibility. The framework remains stable even at extreme values due to the inductive bias and mutual regularization of deep unfolding.

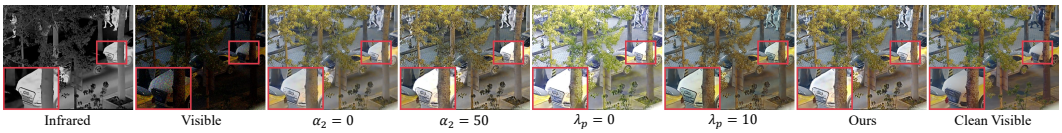


Figure 10: Qualitative results of coefficient analysis.

Quantitatively, as Tab. 15 shows, increasing α_2 improves SCD and Q^{abf} but reduces SSIM and VIF. Removing the coupling term ($\lambda_p = 0$) causes a performance collapse across all metrics. The dynamic

[2]http://www.med.harvard.edu/AANLIB/home.html

strategy $\lambda_p = 2 + 0.1k$ outperforms a fixed large coefficient. The configuration $\alpha_2 = 8$ with dynamic λ_p achieves the best balance, yielding optimal overall performance.

Table 15: Ablation study on hyper-parameters α_2 and λ_p (k: k-th iteration) (best, second-best).

Coefficients	SCD ↑	VIF ↑	Qabf ↑	SSIM ↑
$\alpha_2 = 0, \lambda_p = 2 + 0.1k$	1.350	0.572	0.404	0.836
$\alpha_2 = 50, \lambda_p = 2 + 0.1k$	1.699	0.584	0.512	0.847
$\alpha_2 = 8, \lambda_p = 10$	1.349	0.583	0.462	0.811
$\alpha_2 = 8, \lambda_p = 0$	1.284	0.384	0.368	0.631
$\alpha_2 = 8, \lambda_p = 2 + 0.1k$	1.516	0.591	0.498	0.873

G Theoretical Analysis of the Surrogate Objective

We provide a theoretical analysis of the surrogate objective used in our deep unfolding framework. The following analysis concerns the exact pseudo-inverse that defines the surrogate objective; the learned realization inherits the resulting conditioning and the explicit separation of the null-space component, not the algebraic identities of $D_m^\dagger$. In deep unfolding networks (DUNs), the fidelity objective essentially shapes the optimization landscape, which inherently guides the structural prior and the solution preference of the network. The conventional objective function $\frac{1}{2}\|y_m - D_m z_m\|_F^2$ faces a specific limitation: it does not explicitly constrain the null-space $\mathcal{N}(D_m)$ of the degradation operator. Consequently, its optimization landscape tends to be flat along these null-space directions, lacking a clear preference for recovering unobservable degradation components. This can occasionally lead the iterative solver toward ambiguous solutions. By introducing the surrogate objective $\frac{1}{2}\|D_m^\dagger y_m - z_m\|_F^2$, we can effectively reshape this landscape. We decompose the term inside the surrogate objective as:

$$D_m^\dagger y_m - z_m = D_m^\dagger (y_m - D_m z_m) - (I - D_m^\dagger D_m) z_m. \tag{15}$$

We denote $\mathbf{a} = D_m^\dagger (y_m - D_m z_m)$ and $\mathbf{b} = (I - D_m^\dagger D_m) z_m$. According to the properties of the exact pseudoinverse, $\mathbf{a}$ lies in the row space $\mathcal{R}(D_m^T)$ of the degradation operator D_m, and $\mathbf{b}$ lies in the null space $\mathcal{N}(D_m)$ since $D_m(I - D_m^\dagger D_m) = 0$. Because the row space and null space are orthogonal complements ($\mathcal{R}(D_m^T) \perp \mathcal{N}(D_m)$), $\mathbf{a}$ and $\mathbf{b}$ are strictly orthogonal. Applying the Pythagorean theorem for the Frobenius norm yields:

$$\frac{1}{2}\|D_m^\dagger y_m - z_m\|_F^2 = \frac{1}{2}\|D_m^\dagger (y_m - D_m z_m)\|_F^2 + \frac{1}{2}\|(I - D_m^\dagger D_m) z_m\|_F^2. \tag{16}$$

This decomposition reveals the dual mechanism of the surrogate objective. The first right term, $\frac{1}{2}\|D_m^\dagger (y_m - D_m z_m)\|_F^2$, mitigates the impact of small singular values of D_m, smoothing the narrow valleys into a more well-conditioned landscape. The second term, $\frac{1}{2}\|(I - D_m^\dagger D_m) z_m\|_F^2$, explicitly imposes a null-space regularization. This reshaped landscape induces a mathematical preference for stable, valid solutions, providing a more stable geometric prior for the subsequent task decomposition and encouraging a robust restoration trajectory. Motivated by these profound mathematical properties, our unfolding framework explicitly operates on this surrogate objective, leveraging its robust trajectory to guide the multi-stage iterative solver.

H Limitations

This work handles degradations from both the infrared and visible modality. While the proposed framework effectively addresses common degradations, its robustness might decrease when confronting highly complex or completely unknown hybrid degradations in unpredictable open-world scenarios. Future work will explore restoration techniques for more degradations to extend its applicability to a broader spectrum of extreme degradations.

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 clearly state the proposed deep unfolding framework with contributions containing mutual promoting and interpretability. The claims are supported by the experimental results.

   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 Limitations section in the appendix discusses the limitations of the proposed framework, including its dependence on the assumed degradation model and its behaviour on degradation types not covered by the training setting.

   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: [No]

Justification: The paper provides a theoretical analysis of existing methods and the surrogate objective as a design rationale, but but does not present formal theorems or complete proofs for the proposed network.

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 provide experimental settings including datasets, degradation types, hyperparameter settings and GPUs used for training. The network architecture is comprehensively described in the method section.

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: [No]

   Justification: Code is currently undergoing internal cleanup for better readability and reproducibility. It will be publicly released upon paper acceptance.

   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 paper provide experimental settings including datasets, optimizers, degradation types, hyperparameter settings and GPUs used for training.

   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 single-run results or average values across multiple runs without calculating standard deviations, confidence intervals, or statistical significance tests. No error bars or uncertainty estimates are provided in tables or figures.

   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: [No]

   Justification: The paper reports GPU type and inference time but does not estimate total compute or mention preliminary experiments.

   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: This research uses only publicly available benchmark datasets (LLVIP, MSRS, Harvard) and does not involve potentially harmful applications.

   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: This paper primarily focuses on the fundamental algorithmic design of deep unfolding networks for multi-modal image fusion. This work does not directly deploy or tailor the algorithms for any specific real-world application.

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: This paper does not release any pre-trained models, image generators, or scraped datasets that could be misused. All experiments use standard benchmark datasets which do not contain sensitive or unsafe content.

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: All datasets (LLVIP, MSRS, Harvard) and baseline methods used in this paper are publicly available and properly cited with their original papers.

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: This paper does not introduce any new assets such as datasets, code packages, or pre-trained models.

    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: This paper does not involve crowdsourcing, human subjects, or any data collection that requires participant interaction.

    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: This research does not involve human subjects or crowdsourcing. All experiments use only publicly available benchmark datasets.

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: [No]

Justification: The core methodology of this paper does not involve LLMs as an important, original, or non-standard component. LLMs were used only for minor writing and formatting assistance

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.