Dream-IF: Dynamic Relative EnhAnceMent for Image Fusion

Xingxin Xu[1], Bing Cao[2*], DongDong Li[5], Qinghua Hu[2], Pengfei Zhu[2,3,4*]
[1]School of New Media and Communication, Tianjin University
[2] School of Artificial Intelligence, Tianjin University
[3] Low-Altitude Intelligence Laboratory, Xiong'an National Innovation Center
[4] Xiong'an Guochuang Lantian Technology Co., Ltd.
[5] National Key Laboratory of Science and Technology on Automatic Target Recognition, National University of Defense Technology
{xuxingxin, caobing, huqinghua, zhupengfei}@tju.edu.cn, lidongdong12@nudt.edu.cn

Abstract

Image fusion aims to integrate comprehensive information from images acquired through multiple sources. However, images captured by diverse sensors often encounter various degradations that can negatively affect fusion quality. Traditional fusion methods generally treat image enhancement and fusion as separate processes, overlooking the inherent correlation between them; notably, the dominant regions in one modality of a fused image often indicate areas where the other modality might benefit from enhancement. Inspired by this observation, we introduce the concept of dominant regions for image enhancement and present a Dynamic Relative EnhAnceMent framework for Image Fusion (Dream-IF). This framework quantifies the relative dominance of each modality across different layers and leverages this information to facilitate reciprocal cross-modal enhancement. By integrating the relative dominance derived from image fusion, our approach supports not only image restoration but also a broader range of image enhancement applications. Furthermore, we employ prompt-based encoding to capture degradation-specific details, which dynamically steer the restoration process and promote coordinated enhancement in both multi-modal image fusion and image enhancement scenarios. Extensive experimental results demonstrate that Dream-IF consistently outperforms its counterparts. The code is publicly available.

Introduction

Image fusion aims to integrate essential information from multi-source images captured by various sensors, producing a single comprehensive image. Multi-modal (visible *v.s.* infrared) image fusion (Zhao et al. 2023; Ma et al. 2022; Li and Wu 2018) has been used in a wide range of applications, such as auto driving (Huang et al. 2022), unmanned aerial vehicles (Jasiunas et al. 2002), forest fire monitoring (Liu et al. 2023), etc. Images captured by different imaging sensors often have different characteristics. The infrared images intrinsically avoid visible-light interference but often lack fine texture details. In contrast, the visible images capture abundant color and details in well-lit areas, while may suffer significant quality degradation in complex environments. The fused image preserves the benefits of visible and infrared images while minimizing their limitations by utilizing the complementary information from the respective modalities (Ma, Ma, and Li 2019; Liu et al. 2024).

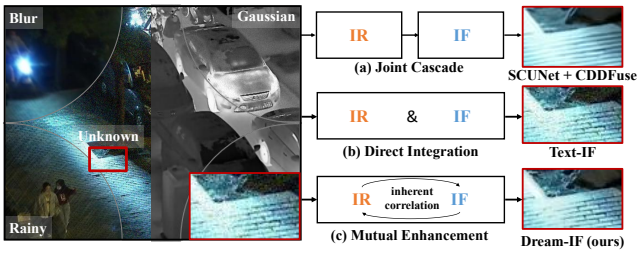

Figure 1: Image fusion (IF) and image restoration (IR) task paradigms. (a) Joint cascaded for restoration followed by fusion, (b) Directly integrate fusion and restoration, (c) Mutually enhanced fusion and restoration through the inherent correlations.

However, due to sensor malfunctions or environmental disturbances, noise is often present in real-world visible and infrared images (Plotz and Roth 2017; Tang, Liu, and Su 2014) resulting in degradation and low quality. Some existing methods (Zamir et al. 2022; Chen et al. 2021) enhance the low-quality images by learning the distribution of datasets and have achieved impressive results. More recently, some approaches (Chen et al. 2022; Wang et al. 2022; Yang et al. 2023) effectively handle degradations by capturing long-range pixel interactions through multi-head attention and feed-forward networks, leveraging the relationship to model the restoration process.

Most existing methods treat image fusion (IF) and image restoration (IR) as two separate tasks (Yang and Li 2009; Xia and Kamel 2007), often requiring restoration to be performed prior to fusion, as illustrated in Fig. 1(a). Recent studies have attempted to address both tasks within a unified framework, leveraging their shared requirement for effective information extraction. CU-Net (Deng and Dragotti 2020) and DeepM^{2}CDL (Deng et al. 2023) designed a unified network to solve general multi-modal image fusion and multi-modal image restoration. It has been demonstrated that multi-modal images can significantly benefit image restoration. Text-IF (Yi et al. 2024) and Text-DiFuse (Zhang, Cao, and Ma 2025) directly merge the two tasks without exploring their mutually reinforcing relationship, potentially leading to suboptimal performance, as illustrated in Fig. 1(b).

*Corresponding author.

Although these works combine image restoration and image fusion in one framework, they still fail to capture their intrinsic coherence.

Intuitively, from the perspective of multi-modal complementarity, image fusion reveals the dominant information of different modalities, which in turn indicates the relative non-dominant information in the other modality, specifically tailoring regions that should be enhanced regardless of degradation issues. By leveraging the relative dominance in image fusion, the difficulty of image degradation recovery can be mitigated. Simultaneously, the use of image degradation tasks can further enhance the adaptability of image fusion to various degradation scenarios.

Based on this motivation, we propose a Dynamic Relative EnhAnceMent framework for Image Fusion (**`Dream-IF`**), jointly boosting the performance of image fusion and enhancement. Unlike most existing methods, Dream-IF first introduces the natural multi-modal complementarity from image fusion to enhance cross-modal image quality, especially for the corresponding non-dominant regions of the respective modality. To perform relative dominance-aware enhancement, we designed the relative enhancement (RE) module to dynamically capture the relative dominance of each modality, which is subsequently employed to facilitate cross-modal enhancement of the relatively weak-quality regions. Consequently, we can obtain the enhanced multi-modal features. By revealing the comprehensive nature of image fusion and converting the dominant regions of one modality to guide the enhancement of the other modality, we enhance the model's capacity for image fusion in complex conditions, making the fusion model more robust to degradations. To the best of our knowledge, this work is the first to explore the intrinsic correlation of image enhancement and fusion using comprehensive relative dominance, rather than treating them as separate data-driven tasks or merely integrating them into a unified framework. Our model copes with robust image fusion for image degradation and even broader image enhancement. The main contributions can be summarized as follows:

- We for the first time investigate the complementarity of image fusion from the perspective of image enhancement. Based on this, we propose a Dynamic Relative EnhAnceMent framework for Image Fusion, termed **`Dream-IF`**, qualified to perform robust image fusion for image degradation and even broader enhancement.
- We design a relative enhancement module, which includes a cross enhancement (CE) module that captures the dynamic relative dominance of each modality to specifically enhance the deficient regions in the other modality, and a self enhancement (SE) module, which further restores the image dynamically using prompts guided by the relative dominance.
- We conduct extensive experiments to evaluate the effectiveness of our framework. Experimental results demonstrate impressive results and generalization, surpassing other competitive methods in both subjective assessments and objective comparisons.

Related work

Image Fusion. Image fusion focuses on extracting effective information from multi-source images and producing an image containing complementary information. Traditional methods often perform image fusion by mathematical transformations, with multiscale analysis (Zhang and Blum 1999), sparse representation (Zhang et al. 2018), subspace learning (Fu et al. 2008), etc. With the development of deep learning, CNN-based methods have made significant advancements. DenseFuse (Li and Wu 2018) pioneered the use of deep learning models for fusing infrared and visible images. Inspired by transform-domain image fusion, IFCNN (Zhang et al. 2020) introduced a general image fusion framework for various fusion tasks. Furthermore, U2Fusion (Xu et al. 2020) addresses various fusion tasks through a unified, unsupervised image fusion network. Similar fusion tasks with aligned objectives can facilitate the integration of complementary information through cross-task commonality. More recently, TC-MoA (Zhu et al. 2024) dynamically selects task-customized mixture of adapters to capture task generality while preserving unique task characteristics. TTD (Cao et al. 2025) proposed an effective test-time fusion paradigm based on generalization theory proving dynamic fusion is superior to static fusion. Considering the potential low-quality of different sensors, some researchers (Deng and Dragotti 2020; Deng et al. 2023) have begun to recognize the correlation between image restoration and fusion. DIVFusion (Tang et al. 2023) enhances low-light images to produce a daytime-like fused image with improved visual perception. Different from these techniques, we first attempt to introduce the naturally complementary dominance of image fusion to dynamically enhance the image quality of non-dominant modality, improving the overall fusion results.

Image Restoration. In recent years, image restoration techniques (Liang et al. 2021; Zamir et al. 2022; Chen et al. 2022; Wang et al. 2022) have achieved significant progress. Particularly, blind restoration (Wu, Dong, and Qiao 2022; Huang and Xia 2020; Soh and Cho 2022) has gained considerable attention. Blind degradation problem is inverse where the degradation is not explicitly known, i.e., non-blind methods can model degradation using prior knowledge (e.g., a predefined blur convolution kernel) while blind methods only assume knowledge of the type of degradation (e.g., blurring)(Chihaoui, Lemkhenter, and Favaro 2024). ECycleGAN (Wu, Dong, and Qiao 2022) tackled the blind reserve problem as an image-to-image task while preserving the fidelity of the reconstructed image. BlindDPS (Chung et al. 2023) solved blind inverse problems by constructing a separate diffusion prior for the forward operator. AirNet (Li et al. 2022) leveraged consistency to learn the degradation representation, recovering various degraded conditions. PromptIR (Potlapalli et al. 2024) introduces a prompt-based learning approach to encode degradation-specific features with prompts and dynamically guide all-in-one image restoration. Despite the extensive exploration of image restoration, these approaches fail to consider the inherent correlation between image restoration and fusion, specifically using the complementarity in image fusion to assist restoration.

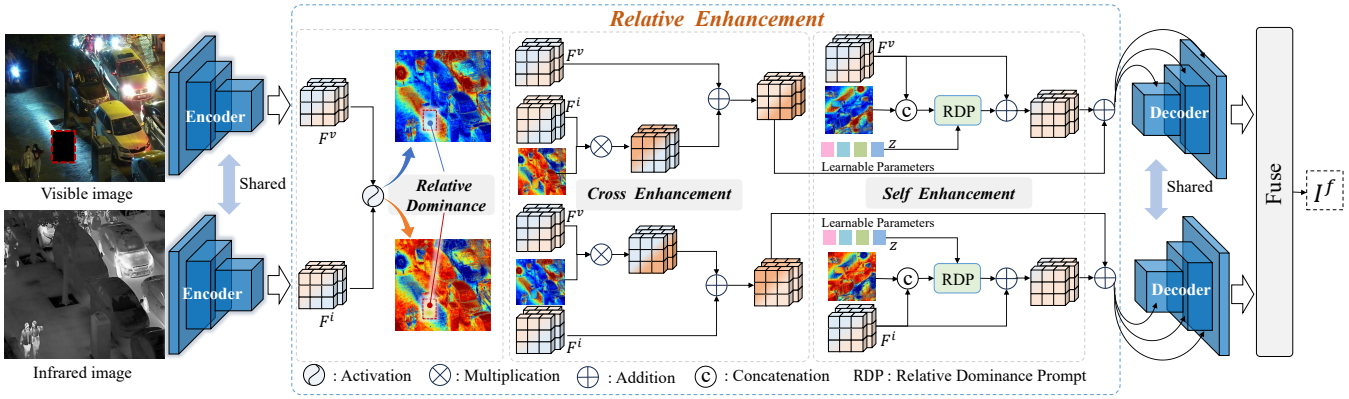

Figure 2: An overview of the proposed Dream-IF. We introduce the Relative Enhancement block, which implicitly enhances non-dominant representations by leveraging relative information during the fusion and restoration process. This module captures the relative dominance inherent in the complementary nature of the image fusion model and uses it to facilitate both cross and self enhancement, ultimately producing the enhanced feature.

Method

We present the pipeline of our Dream-IF in Fig. 2, a dynamic relative enhancement framework acting on the fusion model, which leverages relative dominant prompts to perform cross-modal enhancement for non-dominant regions, thereby facilitating a more robust fusion.

The Overview of Dream-IF

Dream-IF consists of encoders and decoders containing Restormer blocks (Zamir et al. 2022) corresponding to different modalities, which can be defined as $\mathcal{E}$ and $\mathcal{D}$ represent the encoder and decoder, respectively. Given the visible image $I^v \in \mathbb{R}^{H\times W\times 3}$ and the infrared image $I^i \in \mathbb{R}^{H\times W\times 3}$, the feature of each modality can be obtained as follows:

$$F^m = \mathcal{E}(I^m), \tag{1}$$

$$F_i^m = \mathcal{D}_i(F_{i-1}^m), \tag{2}$$

$$I^f = \mathcal{F}(\sum\nolimits_{k\in m} F_N^k) \tag{3}$$

where $F^m \in \mathbb{R}^{h\times w}(m \in \{i, v\})$ represents the latent feature of the given modality, the F_i^m denote the feature at the ith and N is the number of decoder layer. Our goal is to fuse the complementary information from the two source images and obtain a fused image I^f. Thus, the fusion module can be formulated as $\mathcal{F}$, which consists of a Restormer block and a convolution layer to reconstruct the fused image from features.

Relative Enhancement

The encoder and decoder tend to extract contemporary information from source images, which means that when a region of one modality is advantageous, another modality is probably disadvantageous. Inspired by the comprehensive nature of image fusion, we proposed a relative enhancement module to enhance the deficient regions of one modality by capturing the relative dominance of another modality, thus generating a more robust image enhancement and fusion.

Relative Dominance. As discussed above, the image fusion process inherently aims to exploit complementary information from multiple sources. To further enhance this capability, we introduce a mechanism that dynamically preserves complementary features while suppressing redundant information. Specifically, we define the fusion weight for each source feature as its Relative Dominance (RD), formulated as $\mathrm{RD}^m = \sigma\mathrm{Conv}(F^m)$, where σ denotes an activation function. The optimization objective is defined as:

$$\min \mathbb{E}[\log(1 - \sum_{h,w}^{N} \mathrm{RD}^{(m)})]. \tag{4}$$

The fusion process aims at capturing the dominance of each source and leveraging it to enhance non-dominant areas. The visualizations of RD are shown in the Appendix.

Relative Enhancement. In image fusion, the dominance of one modality can be approximated by the non-dominance of another modality, which should be appropriately enhanced. Thus, the primary goal of the relative enhancement (RE) module is to boost the representation of non-dominant features by facilitating the relative dominance of another modality.

To naturally integrate the RD and the feature to be strengthened, we present the cross enhancement (CE) and self enhancement (SE) module to enhance. The RD denotes the relative dominance from the other modality and indicates its deficient parts to be enhanced, thereby efficiently guiding the reconstruction process. Specifically, the CE module addresses deficiencies in one modality by complementing it with information from the other modality. It leverages the complementary nature of the modalities, filling in the gaps of the defective feature with the more dominant or reliable feature from the other modality. This ensures a more complete and accurate representation of the scene. The restoration process is guided by the dominant region in the other modality, where the strength of one modality helps identify areas in the other that require enhancement. By focusing on these dominant regions, the module effectively restores the

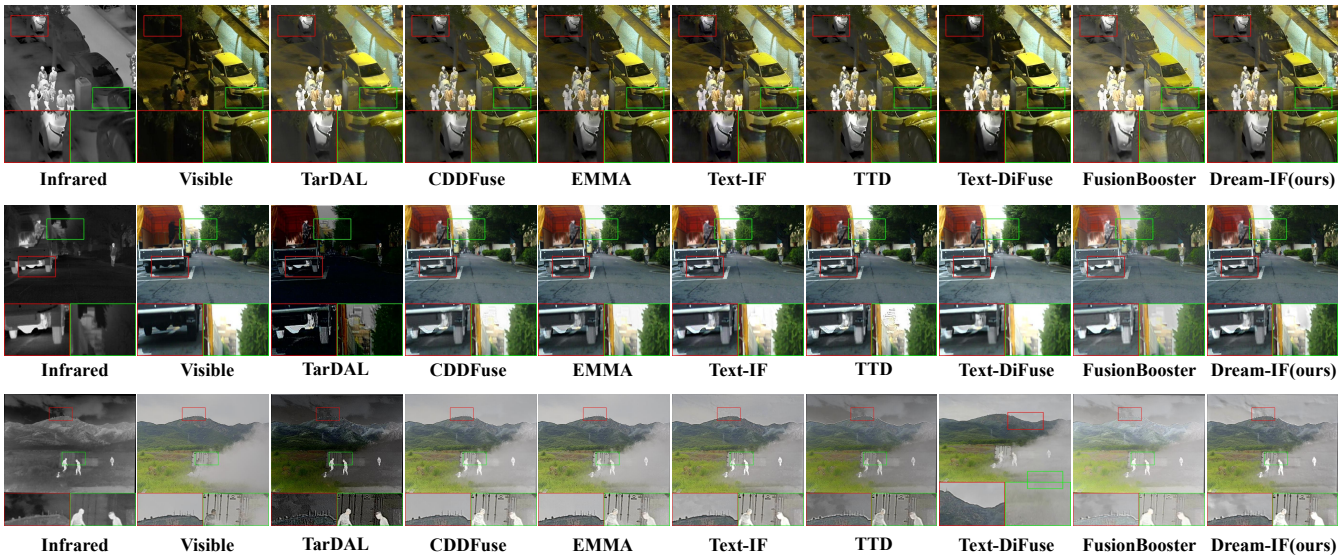

Figure 3: Qualitative comparisons of various methods on representative images selected from the LLVIP, MSRS and M3FD datasets.

missing or degraded information in the subordinate modality, resulting in a more coherent and refined output. It can be formulated as

$$\dot{F}^m = F^m + \mathbf{CE}(F^m, RD). \tag{5}$$

Further to enhance itself and restore degeneration, to make our model fundamentally resistant to degradation (Potlapalli et al. 2024), we generate a Self Enhancement (SE). This is calculated as:

$$\hat{F}^m = F^m + \mathbf{SE}(\mathbf{RDP}(F^m, RD, z), F^m) \tag{6}$$

where z is a set of learnable parameters. The Relative Dominance Prompt (RDP) serves as an adaptive enhancement module, which generates restoration prompts dynamically based on the input feature F^m and its corresponding relative dominance RD. Further architectural details are provided in the Appendix. Afterward, the refined feature is sent to the next block in the decoder, where it is transformed into the denoised and restored feature:

$$F_l^m = \mathcal{D}_l(\widetilde{F}_{l-1}^m), \tag{7}$$

where $\widetilde{F} = \text{Conv}(\mathbf{Cat}[\dot{F}; \hat{F}])$. This enhanced feature is then passed to the fusion block, which combines the features from the two modalities and generates the final fused image.

Loss Function

Dream-IF trains with Gaussian noise degradation, add in clean data. The loss functions we used, including pixel loss, gradient loss, SSIM loss, and color loss, which can be depicted as:

$$\mathcal{L}_f = \mathcal{L}_{pixel} + \mathcal{L}_{grad} + \mathcal{L}_{SSIM} + \mathcal{L}_{color}. \tag{8}$$

Moreover, to explicitly optimize the relative dominance (RD), we define the following loss function:

$$\mathcal{L}_{RD} = \sum_j^h \sum_k^w \sum_m^N RD_{i,j}^m - 1. \tag{9}$$

The overall loss function is defined as $\mathcal{L} = \mathcal{L}_f + \mathcal{L}_{\text{RD}}$.

Experiments

In this section, We conduct comprehensive qualitative and quantitative comparisons to evaluate the performance of our proposed method against state-of-the-art approaches. To further validate the contribution of each component, we perform an extensive ablation study.

Experimental Setting

Implementation Details. Our experimental framework is implemented using PyTorch (Paszke et al. 2019) and executed on an NVIDIA 3090 GPU. The proposed architecture adopts a U-Net-based design with a symmetric 4-level encoder-decoder structure, where the encoder extracts hierarchical features and the decoder reconstructs the output image. Each level of the encoder and decoder incorporates multiple Restormer (Zamir et al. 2022) blocks, with the number of blocks progressively increasing from higher to lower levels. The complete architectural details, including layer configurations and hyperparameters, are provided in the supplementary material for reproducibility.

Datasets. Our experiments are conducted on three widely recognized publicly available datasets: LLVIP, MSRS, and M3FD. Further details about the datasets are provided in the supplement. Following (Zhu et al. 2024), utilizing $12,025$ image pairs from the LLVIP dataset with random degradation for training and 70 image pairs for testing. To rigorously assess the generalization capability of the proposed method, we tested the model on both the MSRS and M3FD datasets without any fine-tuning.

Competing Methods. We compared our approach with seven recent competing methods, including TarDAL (Liu et al. 2022), CDDFuse (Zhao et al. 2023), EMMA (Zhao et al. 2024), Text-IF (Yi et al. 2024), TTD (Cao et al. 2025), Text-DiFuse (Zhang, Cao, and Ma 2025), and FusionBooster (Cheng et al. 2025).

Method	LLVIP Dataset					MSRS Dataset					M3FD Dataset				
	AG	SF	Q^{abf}	SSIM	VIFF	AG	SF	Q^{abf}	SSIM	VIFF	AG	SF	Q^{abf}	SSIM	VIFF
TarDAL	4.921	18.207	0.410	1.080	0.537	1.914	5.944	0.170	0.930	0.380	4.331	15.766	0.292	0.731	0.486
CDDFuse	5.403	18.495	0.582	1.184	0.660	3.779	11.570	0.585	1.303	0.717	4.800	14.709	0.521	1.379	0.578
EMMA	5.560	17.190	0.552	1.192	0.648	3.775	11.559	0.621	1.324	0.738	**5.362**	15.304	0.590	1.377	**0.711**
Text-IF	5.682	17.718	0.640	1.074	**0.813**	3.801	11.868	0.624	1.117	**0.867**	5.034	15.494	**0.644**	1.339	0.637
TTD	5.509	19.914	0.627	1.196	0.730	3.705	11.527	0.549	1.349	0.656	5.022	15.254	0.519	1.367	0.607
Text-Dif	4.848	15.528	0.403	1.045	0.524	3.842	11.506	0.436	0.935	0.703	2.810	8.465	0.149	1.122	0.165
FusionBooster	5.476	15.464	0.407	0.978	0.551	3.208	9.012	0.420	1.043	0.630	3.559	10.245	0.396	1.297	0.515
Dream-IF	**5.926**	**19.992**	**0.679**	**1.198**	0.800	**3.962**	**12.293**	**0.631**	**1.362**	0.849	5.124	**16.198**	0.614	**1.387**	0.704

Table 1: Quantitative comparison with SOTAs without degradation. The **bold**/underline indicates the best and runner-up.

Evaluation Metrics. We evaluated the fusion results quantitatively on five metrics, including the average gradient (AG), spatial frequency (SF), gradient-based similarity measurement Q^{abf} (Piella and Heijmans 2003), structural similarity (SSIM) (Wang et al. 2004), and visual information fidelity for fusion (VIFF) (Han et al. 2013).

Comparison without Degradation

We conduct comprehensive qualitative and quantitative evaluations to compare the performance of the proposed Dream-IF against state-of-the-art competing methods under the scenario without degradation.

Qualitative Comparison. Our comprehensive evaluation reveals that, while all competing fusion methods are capable of combining primary features from infrared and visible images to some extent, the proposed method demonstrates clear advantages that significantly enhance fusion quality. Our method excels in emphasizing the critical characteristics of infrared images while preserving complementary visible information. As illustrated in Fig. 3 (LLVIP dataset, red boxes), the visible images in low-light conditions are often blurred, whereas the infrared images provide essential complementary details. Competing methods, however, suffer from significant infrared content loss in the fusion results, as evidenced by the dark and indistinct details of the cars in the green boxes. Although CDDFuse shows competitive visual performance in the red boxes, it fails to retain sufficient infrared information, leading to suboptimal fusion in the green boxes. In contrast, our method enhances vulnerable modalities through a robust feature integration mechanism, effectively preserving both infrared and visible information. This results in superior perceptual quality and improved information retention, as demonstrated by the clear and detailed fusion results in both highlighted regions. The robust experimental results show consistent performance on both the MSRS and M3FD datasets.

Quantitative Comparison. The quantitative evaluation, conducted using five widely recognized metrics across the LLVIP, MSRS, and M3FD datasets, is summarized in Table 1. On the LLVIP dataset, our method achieves the highest scores across most metrics, namely AG, SF, Q^{abf}, and SSIM. Notably, the highest AG and SF scores indicate that our method retains the richest information and sharpest details, attributed to its ability to enhance non-dominant regions while preserving dominant features dynamically. The superior Q_{abg} score reflects better alignment of local gradients and intensities between the source and fused images, demonstrating our method's capability to effectively integrate multi-modal information. Furthermore, the highest SSIM score confirms that the fused images retain a substantial amount of visual information from the source modalities, underscoring the high perceptual quality of our results. These quantitative findings align with the qualitative observations, validating that our method achieves superior fusion performance through dynamic relative enhancement and robust multi-modal integration. Our method also demonstrates competitive performance across most evaluation metrics on the MSRS and M3FD datasets. The consistency across datasets further validates the generalizability and robustness of our approach under diverse scenarios.

Overall, the quantitative results reinforce the effectiveness of our method in achieving high-quality image fusion, as evidenced by its ability to retain rich information, preserve structural details, and align multi-modal features effectively.

Methods	AG	SF	Q^{abf}	SSIM	VIFF
TarDAL	4.745	17.337	0.266	0.503	0.256
CDDFuse	4.530	15.929	0.523	1.179	0.618
EMMA	4.593	14.818	0.499	1.186	0.592
Text-IF	4.863	16.554	0.536	1.156	0.609
TTD	4.948	16.637	0.489	1.144	0.533
Text-DiFuse	4.084	13.777	0.364	1.049	0.490
FusionBooster	4.524	13.401	0.380	1.017	0.533
Dream-IF	**5.243**	**17.441**	**0.580**	**1.186**	**0.705**

Table 2: Comparison with SOTAs under degradation. The **bold** indicates the best.

Comparison with Degradation

To assess the robustness and real-world applicability of our method, we introduce a diverse set of degradations (Zhang et al. 2021) to simulate challenging scenarios that amplify modality discrepancies. These degradations include Gaussian noise, Poisson noise, speckle noise, and other common artifacts encountered in practical applications (e.g., blur, etc.). Additionally, we incorporate synthetic rainy degradations to further mimic real-world conditions. For a fair comparison, we compare existing methods for image fusion and

restoration. This includes two-stage fusion, which performs degradation recovery followed by fusion, and one-stage fusion with degradation.

Degradations. Due to the inherent limitations of imaging conditions, various types of noise inevitably arise in practical scenarios, significantly impacting the quality and usability of captured images. To accurately simulate real-world degradation, we model three fundamental types of noise based on their physical origins in optical imaging systems.

Gaussian Noise primarily results from inherent electronic noise in image acquisition or transmission devices, as well as signal interference in low signal-to-noise ratio (SNR) environments. Mathematically, Gaussian noise can be expressed as: $y = x + \mathcal{N}(0, \sigma)$. Poisson noise, also referred to as shot noise, arises from the inherent quantum uncertainty in photon counting during the light detection process, which can be formulated as: $y = x + \frac{\lambda^k e^{-\lambda}}{k!}$. Speckle noise is caused by random phase interference in coherent imaging processes, representing structural corruption rather than additive distortion. It is modeled as: $y = x \odot (1 + \mathcal{N}(0, \epsilon)$. In this paper, we set $\sigma = 35$, $\lambda \in [2, 4]$, and $\epsilon \in [2, 25]$.

Each training sample undergoes one or more random degradations selected from the aforementioned types. Current image fusion methods often struggle with blind degradation, frequently leading to suboptimal fusion results. Our approach addresses this limitation by narrowing the domain gap and applying cross-modal enhancement for restoration to improve the final output.

Two-stage Fusion. For blind restoration, we apply random degradation to the LLVIP test set. To ensure a fair comparison, we employ a two-stage fusion strategy. Specifically, for all the competing methods, we adopt SCUNet (Zhang et al. 2023), a widely recognized restoration model, to restore images before fusion. For comparison, our Dream-IF directly performs image restoration and fusion in a unified model, which is a more challenging.

Quantitative results. Our method compared with the combined results of SOTA image fusion methods after SCUNet restoration on degraded source images are presented in Fig. 4. Our method retains more details, exhibiting a more realistic overall texture, particularly in the green box. CDDFuse exhibits competitive performance but also suffers from detail and texture loss. TarDAL suffers from fidelity loss, while Text-IF, TTD, Text-DiFuse and FusionBooster produce excessively smooth results, omitting important information. Although EMMA performs competitively in the green box, it fails to preserve details in the red box as effectively as ours. The advantages demonstrate that our approach effectively performs both restoration and fusion, with its strengths becoming more pronounced as the differences between modalities increase.

Qualitative results are reported in Table 2, demonstrating that our method achieves competitive results even without the use of a specialized restoration model. Specifically, our method achieves optimal values across all metrics. The highest AG and SF scores indicate that Dream-IF excels at retaining texture and detail. VIFF, SSIM, and Q^{abf} indicate that our method effectively preserves mutual information between the source images, ensuring details and contextual information are retained in the output while maintaining subjective visual quality. It underscores the ability of our approach to preserve image fidelity and enhance degraded features without relying on a dedicated restoration model.

Methods	Rain	Gaussian	Poisson	Speckle	Blur	Resize
Text-IF*	0.273	0.275	0.275	0.270	0.269	0.244
Text-DiFuse	0.296	0.291	0.268	0.266	0.290	0.268
Dream-IF	**0.241**	**0.255**	**0.242**	**0.240**	**0.260**	**0.241**

Table 3: One-stage fusion comparison. **Bold** is the best. Text-IF* is Text-IF with description text.

One-stage Fusion. To verify the performance of our method in simultaneous fusion and restoration within a single framework, we further evaluate our Dream-IF with Text-IF, a text-guided restoration and fusion model. Since Text-IF requires descriptive text to specify the degradation type, we provide degradation text guidance for Text-IF. It is worth noting that our model does not take any prior knowledge of the degradation type, which is a much more challenging setting to handle blind restoration and fusion. We evaluated image restoration using the learned perceptual image patch similarity (LPIPS). LPIPS captures perceptual differences between images, with lower scores indicating greater similarity.

Qualitative results. As presented in Fig. 5, it demonstrates that our method recovers more texture details affected by degradation. However, even with degradation descriptions, Text-IF fails to restore images affected by Gaussian noise fully, and it results in overly smooth fused images when applied to other types of noise. Text-DiFuse removes noise through diffusion, but yields poor performance. This demonstrates that our method is robust to various degradations, benefiting from dynamic dominance enhancement to improve inferior features. In contrast, our method leverages information from both modalities to seamlessly integrate image fusion and restoration, enabling effective blind restoration and fusion. It further validates our effectiveness in utilizing the comprehensive nature of fusion to drive overall image enhancement in the multi-modal fusion task.

Quantitative results. As shown in Table 3, our method consistently outperforms in terms of LPIPS scores across most degradation types. Notably, the method performs exceptionally well when applied to images affected by Gaussian noise. Gaussian noise is particularly challenging because it introduces subtle distortions that affect all frequencies within the image. These distortions are difficult to mitigate without compromising finer image details or introducing artifacts. The lower LPIPS scores we achieve in this context are indicative of higher perceptual similarity, meaning that our approach is better at maintaining the natural textures and structures of the image. This ability to preserve details without significant loss of fidelity means that our framework produces results that are visually closer to the original, real image, making it a highly effective solution for addressing Gaussian noise degradation. Furthermore, the preservation of fine-grained textures and intricate features enhances the

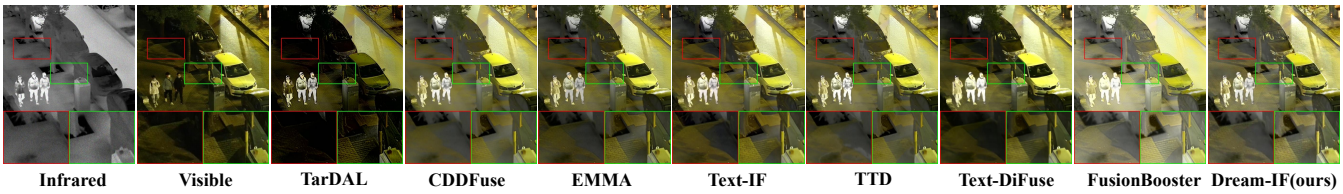

Figure 4: Qualitative comparisons of various methods under degradation on the LLVIP dataset. SCUNet+ refers to the application of the SCUNet model for image restoration prior to the fusion process.

Base	RD	CE	SE	EI	AG	PSNR	Q^{abf}	VIFF
✓	-	-	-	5.402	18.032	0.627	1.181	0.736
✓	✓	-	-	5.638	18.594	0.659	1.185	0.756
✓	✓	✓	-	5.642	18.633	0.675	1.193	0.770
✓	✓	-	✓	5.746	19.446	0.668	1.191	0.797
✓	✓	✓	✓	**5.926**	**19.992**	**0.679**	**1.198**	**0.800**

Table 4: Ablation studies on LLVIP dataset. **Bold** is the best.

Methods	Recall	mAP@.5	mAP@.5:.95
TarDAL	0.8850	0.9495	0.6291
CDDFuse	0.8927	0.9496	0.6031
EMMA	0.8894	0.9502	0.6192
Text-IF	0.8951	0.9517	0.6148
TTD	0.8888	0.9497	0.6252
Text-DiFuse	0.8742	0.9394	0.5964
FusionBooster	0.8925	0.9502	0.6245
Dream-IF	**0.9041**	**0.9543**	**0.6327**

Table 5: Comparison of object detection on LLVIP dataset. **Blod** indicates the best.

overall quality of the images processed by our method, distinguishing it from other approaches that might smooth or blur these details.

Performance on Downstream Task

We perform experiments on object detection to verify the compatibility of Dream-IF with high-level downstream tasks. Specifically, we train the YOLOv11 (Khanam and Hussain 2024) detector on the LLVIP dataset, using fusion results generated by competitive comparison methods, and evaluate performance using recall and mAP. As shown in Table 5, Dream-IF outperforms the competition in object detection, achieving the best performance in recall, mAP@.5, and mAP.5:.95. Additionally, visualizations of the comparisons in object detection are shown in the Appendix.

Ablation Study

We conduct ablation studies on the LLVIP dataset, as shown in Table 4, to validate the effectiveness of the proposed CE and SE module. Specifically, we evaluate the performance of our model by removing the CE and SE modules, treating the model without these components as the baseline. The

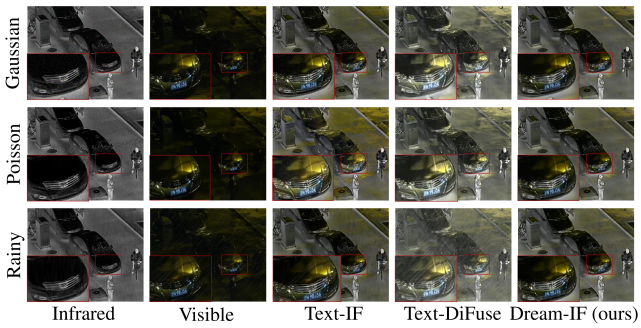

Figure 5: Visualization of comparison in one-stage restoration.

results demonstrate that adding either the CE or SE module leads to minor improvements in the performance across all evaluation metrics. When both modules are incorporated into the model, substantial improvements in all indicators are observed. These results highlight the distinct and complementary contributions of each module. Notably, the inclusion of both CE and SE along with the final model configuration leads to the highest scores across all metrics, confirming that our method achieves the best qualitative and quantitative performance among all ablation settings. This further validates the importance of the CE and SE modules, as well as their combined effect in enhancing the overall model's capabilities. The improvements across various indicators demonstrate the efficacy of our approach in achieving superior results.

Conclusion

In this paper, we first explore the inherent connection between image fusion and image enhancement. Based on the observation that the dominant region of one modality in image fusion relatively indicates the inferiors of the other modality, we propose a dynamic relative enhancement framework for image fusion (Dream-IF). We extract the relative dominance from the fusion model for each sample and use it to perform prompting cross-modal enhancement. Notably, our Dream-IF is a blind restoration model, coping with robust image fusion for image degradation and even broader image enhancement. Extensive experiments validate our effectiveness against the competing methods. In the future, we will further study the potential of relative dominance derived from image fusion in other related tasks.

Acknowledgments

This work was sponsored by the National Natural Science Foundation of China (No.s 62222608, 62436002, 62476198), the Tianjin Natural Science Funds for Distinguished YoungScholar (No. 23JCJQJC00270), the Zhejiang Provincial Natural Science Foundation of China (No. LD24F020004), and the Natural Science Foundation of Tianjin (No. 25JCYBJC00950).

References

Cao, B.; Xia, Y.; Ding, Y.; Zhang, C.; and Hu, Q. 2025. Test-Time Dynamic Image Fusion. *Advances in Neural Information Processing Systems*, 37: 2080–2105.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *European conference on computer vision*, 17–33. Springer.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 182–192.

Cheng, C.; Xu, T.; Wu, X.-J.; Li, H.; Li, X.; and Kittler, J. 2025. Fusionbooster: A unified image fusion boosting paradigm. *International Journal of Computer Vision*, 133(5): 3041–3058.

Chihaoui, H.; Lemkhenter, A.; and Favaro, P. 2024. Blind Image Restoration via Fast Diffusion Inversion. *arXiv preprint arXiv:2405.19572*.

Chung, H.; Kim, J.; Kim, S.; and Ye, J. C. 2023. Parallel Diffusion Models of Operator and Image for Blind Inverse Problems. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Deng, X.; and Dragotti, P. L. 2020. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3333–3348.

Deng, X.; Xu, J.; Gao, F.; Sun, X.; and Xu, M. 2023. DeepM 2 CDL: Deep Multi-scale Multi-modal Convolutional Dictionary Learning Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Fu, Y.; Cao, L.; Guo, G.; and Huang, T. S. 2008. Multiple feature fusion by subspace learning. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, 127–134.

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

Huang, K.; Shi, B.; Li, X.; Li, X.; Huang, S.; and Li, Y. 2022. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*.

Huang, L.; and Xia, Y. 2020. Joint blur kernel estimation and CNN for blind image restoration. *Neurocomputing*, 396: 324–345.

Jasiunas, M. D.; Kearney, D. A.; Hopf, J.; and Wigley, G. B. 2002. Image fusion for uninhabited airborne vehicles. In *2002 IEEE International Conference on Field-Programmable Technology, 2002.(FPT). Proceedings.*, 348–351. IEEE.

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

Li, B.; Liu, X.; Hu, P.; Wu, Z.; Lv, J.; and Peng, X. 2022. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17452–17462.

Li, H.; and Wu, X.-J. 2018. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liu, J.; Fan, X.; Huang, Z.; Wu, G.; Liu, R.; Zhong, W.; and Luo, Z. 2022. 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*, 5802–5811.

Liu, J.; Wu, G.; Liu, Z.; Wang, D.; Jiang, Z.; Ma, L.; Zhong, W.; and Fan, X. 2024. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Y.; Zheng, C.; Liu, X.; Tian, Y.; Zhang, J.; and Cui, W. 2023. Forest Fire Monitoring Method Based on UAV Visual and Infrared Image Fusion. *Remote Sensing*, 15(12): 3173.

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

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Piella, G.; and Heijmans, H. 2003. A new quality metric for image fusion. In *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, volume 3, III–173. IEEE.

Plotz, T.; and Roth, S. 2017. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1586–1595.

Potlapalli, V.; Zamir, S. W.; Khan, S. H.; and Shahbaz Khan, F. 2024. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36.

Soh, J. W.; and Cho, N. I. 2022. Variational deep image restoration. *IEEE Transactions on Image Processing*, 31: 4363–4376.

Tang, L.; Liu, L.; and Su, J. 2014. Modeling and simulation research of infrared image noise. *Infrared Technol*, 36: 542–548.

Tang, L.; Xiang, X.; Zhang, H.; Gong, M.; and Ma, J. 2023. DIVFusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91: 477–493.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17683–17693.

Wu, S.; Dong, C.; and Qiao, Y. 2022. Blind image restoration based on cycle-consistent network. *IEEE Transactions on Multimedia*, 25: 1111–1124.

Xia, Y.; and Kamel, M. S. 2007. Novel cooperative neural fusion algorithms for image restoration and image fusion. *IEEE transactions on image processing*, 16(2): 367–381.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2020. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518.

Yang, B.; and Li, S. 2009. Multifocus image fusion and restoration with sparse representation. *IEEE transactions on Instrumentation and Measurement*, 59(4): 884–892.

Yang, Y.; Cao, S.; Wan, W.; and Huang, S. 2023. Multimodal medical image super-resolution fusion based on detail enhancement and weighted local energy deviation. *Biomedical Signal Processing and Control*, 80: 104387.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2024. 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*, 27026–27035.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhang, H.; Cao, L.; and Ma, J. 2025. Text-DiFuse: An Interactive Multi-Modal Image Fusion Framework based on Text-modulated Diffusion Model. *Advances in Neural Information Processing Systems*, 37: 39552–39572.

Zhang, K.; Li, Y.; Liang, J.; Cao, J.; Zhang, Y.; Tang, H.; Fan, D.-P.; Timofte, R.; and Gool, L. V. 2023. Practical blind image denoising via Swin-Conv-UNet and data synthesis. *Machine Intelligence Research*, 20(6): 822–836.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4791–4800.

Zhang, Q.; Liu, Y.; Blum, R. S.; Han, J.; and Tao, D. 2018. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Information Fusion*, 40: 57–75.

Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; and Zhang, L. 2020. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118.

Zhang, Z.; and Blum, R. S. 1999. A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. *Proceedings of the IEEE*, 87(8): 1315–1326.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023. 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*, 5906–5916.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Zhang, K.; Xu, S.; Chen, D.; Timofte, R.; and Van Gool, L. 2024. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25912–25921.

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