Unified Text-Guided Degradation-Aware Fusion of Infrared and Visible Images

1st Yueqiang Zhao
Institute of Software, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
zhaoyueqiang2024@iscas.ac.cn

2nd Yijun Lin
Institute of Software, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
yijun2016@iscas.ac.cn

3rd Xiongxin Tang*
Institute of Software, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
xiongxin@iscas.ac.cn

***Abstract*—Existing image fusion methods often perform poorly in real-world scenarios with complex degradations due to a lack of understanding of diverse degradation characteristics and the absence of relevant datasets. To address this issue, we propose a unified text-guided degradation-aware fusion framework that introduces textual prompts to adaptively mitigate multiple degradations. A two-stage training strategy is adopted, together with a Cross-Prompt Attention Aggregation Module and a JointGate module, to enable cross-degradation knowledge sharing. In addition, we construct an infrared-visible fusion dataset that covers various representative degradation scenarios, providing a solid benchmark for evaluating fusion performance under complex conditions and facilitating future research on degradation-aware image fusion. Extensive experiments and ablation studies demonstrate that the proposed method achieves superior performance in both single and compound degradation cases, showing stronger generalization capability. Our code will be available comming soon.**

***Index Terms*—Image fusion, Infrared and visible, Degradation scenarios, Text guidance**

I. Introduction

Infrared and visible image fusion (IVIF) aims to generate images that retain salient infrared targets and fine visible details. Such fused images can enhance downstream tasks, including detection, segmentation, and re-identification. However, in real-world scenarios, IVIF faces multiple degradations that compromise quality. Visible images are prone to low light or overexposure, while infrared images suffer from noise and low contrast. Such degradations reduce clarity and hinder downstream performance.

Current methods follow two paradigms. The first is a two-stage approach: restoration models first remove degradations before fusion. However, due to the significant domain gap between restoration and fusion models, this strategy is susceptible to domain shift issues, making it difficult to guarantee the quality of the final fused images. The second paradigm focuses on degradation-specific designs, such as DIVFusion [1] based on Retinex theory, or PFF [2] guided by visual perception. While effective for single degradations, these methods struggle with complex combinations. In particular, when visible and infrared images suffer from different degradations (e.g., low-light in visible images combined with low contrast in infrared images), designing dedicated fusion models for each case is costly and impractical. Nevertheless, robust IVIF in all-weather and complex scenarios requires a fusion model with adaptive degradation-handling capability.

With the rise of vision-language models, text prompts offer flexible guidance for image processing. On the one hand, semantic text provides high-level guidance, enabling personalized and controllable enhancement. On the other hand, the strong generalization capability of pretrained language models helps overcome the limitations of degradation-specific priors. Text-IF [3] was the first to introduce text guidance into the IVIF task, achieving degradation-aware and interactive fusion through semantic prompts. However, it relies on fixed prompts, lacking adaptability to compound degradations.

To address these challenges, we propose a unified text-guided degradation-aware IVIF framework. Unlike existing approaches, our method adopts **two-stage training paradigm**. In the first stage, the model learns stable fusion under individual degradations. We employ **Semantic-Guided Cross-Attention Module**, which leverages the attention mechanism to enable text prompts to spatially regulate the fusion of infrared and visible features, enhancing controllability and robustness. In the second stage, semantic alignment and knowledge sharing across tasks are achieved through **Cross-Prompt Attention Aggregation Module** and **JointGate Module**, producing unified text-prompt representations for diverse degradations. This is the first attempt in the field of visible light and infrared image fusion. Our method avoids the high computational cost of training separate models for each degradation combination.

*Corresponding author: Xiongxin Tang, xiongxin@iscas.ac.cn.

This work is supported by the National Key Laboratory of Space Integrated Information System under Grant No. E5GF5910.

In addition, we have constructed a **multi-degradation infrared-visible fusion dataset**, which encompasses both single-degradation and compound-degradation scenarios. We designed dedicated training and testing sets for each degradation scenario, offering a comprehensive and challenging benchmark that facilitates the development and evaluation of robust multi-degradation perception fusion methods.

The main contributions of this work are as follows:

- A unified text-guided fusion framework with two-stage training, achieving both stable single-degradation performance and adaptive multi-degradation fusion.
- A Cross-Prompt Attention Aggregation Module and JointGate Module that align semantics across degradations to learn unified prompts.
- A Semantic-Guided Cross-Attention module that spatially regulates fusion with text, improving controllability and robustness.
- A multi-degradation dataset covering typical scenarios, providing a benchmark for future research.

II. Related Work

Image Fusion. Earlier visual-oriented fusion approaches primarily concentrated on merging complementary information from multiple modalities and improving visual fidelity. These methods often depend on sophisticated network architectures and carefully designed loss functions to ensure that the integrated content retains the essential characteristics of the original images, preserving their unique features while enhancing overall quality. The mainstream network architectures primarily include CNNbased [4], [5], AE-based [6], [7], GAN-based [8], Transformers [9] and diffusion models [10]. Furthermore, several schemes including joint registration and fusion [11], [12], semantic-driven [13], [14], and degradationrobust [3] methods, are proposed to broaden the practical applications of image fusion. For instance, Tang et al. [15] and Liu et al. [16] developed corresponding solutions for illumination distortions and noise interference, respectively. In addition, Yi et al. [3] proposed a unified image restoration and fusion network to address various degradations. However, they are unable to handle mixed degradations and struggle to generalize to real-world scenarios. In particular, Text-IF [3] relies on tedious manual efforts to customize text prompts for each scene, hindering large-scale automated deployment.

Image Restoration with Vision-Language Models. With the advancement of deep learning toward multimodal integration, the image restoration field has progressively evolved into a new paradigm of text-driven schemes. CLIP [17] establishes visual-textual semantic representation alignment through dual-stream Transformer encoders pre-trained on 400 million image-text pairs. This framework lays the foundation for the widespread adoption of Prompt Engineering in computer vision. Recent studies integrate CLIP encoders with textual user instructions, proposing generalized restoration frameworks, including PromptIR [18], AutoDIR [19], and InstructIR [20], which effectively handle diverse degradation types. Moreover, SPIRE [21] further introduces fine-grained textual restoration cues by quantifying degradation severity levels and supplementing semantic information for precise restoration. To eliminate reliance on manual guidance, Luo et al. [22] developed DA-CLIP by fine-tuning CLIP on mixed degradation datasets, enabling degradation-aware embeddings to facilitate distortion handling. However, existing unified restoration methods are primarily designed for natural images, which exhibit notable limitations when processing multimodal image pairs (e.g., infrared-visible).

III. Proposed Method

A. Problem Analysis

Given two source images, an infrared image $I_{ir} \in \mathbb{R}^{H\times W\times 1}$ and a visible image $I_{vi} \in \mathbb{R}^{H\times W\times 3}$, a typical fusion paradigm employs a deep network $\mathcal{N}_f$ to integrate complementary information and generate a fused image I_f, formulated as:

$$I_f = \mathcal{N}_f(I_{ir}, I_{vi}). \tag{1}$$

In real-world scenarios, the source images are often degraded by various factors, which severely deteriorate the quality of fusion. Therefore, advanced fusion paradigms must simultaneously consider image restoration and information integration. To address this challenge, we introduce textual prompts into the fusion process, breaking the limitation of producing a single fused result while enhancing robustness against degradations. Specifically, we adopt a two-stage training strategy to establish a unified text-guided degradation-aware framework.

In the first stage, we use infrared-visible image pairs with single-type degradations, along with corresponding textual prompts, to guide the model in learning basic degradation-aware enhancement and fusion capabilities. This process can be expressed as:

$$I_f = \mathcal{N}_{rf}(I_{ir}, I_{vi}, T_{text} \mid \Omega), \tag{2}$$

where T_{text} denotes the textual prompt for a specific degradation, and Ω represents the set of degradation types.

In the second stage, when the model is exposed to images suffering from multiple degradations, it receives multiple textual prompts and dynamically fuses them to adaptively align semantic information across tasks, thereby achieving unified cross-degradation modeling. This can be formulated as:

$$I_f = \mathcal{N}_{rf}\big(I_{ir}, I_{vi}, \mathcal{G}(T_{text}^{(1)}, T_{text}^{(2)}, \ldots, T_{text}^{(N)}) \mid \Omega\big), \tag{3}$$

where $\mathcal{G}$ denotes the cross-task prompt alignment function.

B. Image Fusion Pipeline

Image Encoder. The encoder extracts multi-scale features from degraded infrared (D_{ir}) and visible (D_{vi}) images using Transformer/Restormer-based blocks, which capture both spatial and semantic information.

Cross-Fusion Layer. To integrate features across modalities, we adopt cross-attention (CR-ATT). Each modality is projected into Q, K, and V.

$$\{Q_{ir}, K_{ir}, V_{ir}\} = \mathcal{F}_{ir}^{qkv}(F_{ir}), \quad \{Q_{vi}, K_{vi}, V_{vi}\} = \mathcal{F}_{vi}^{qkv}(F_{vi}), \tag{4}$$

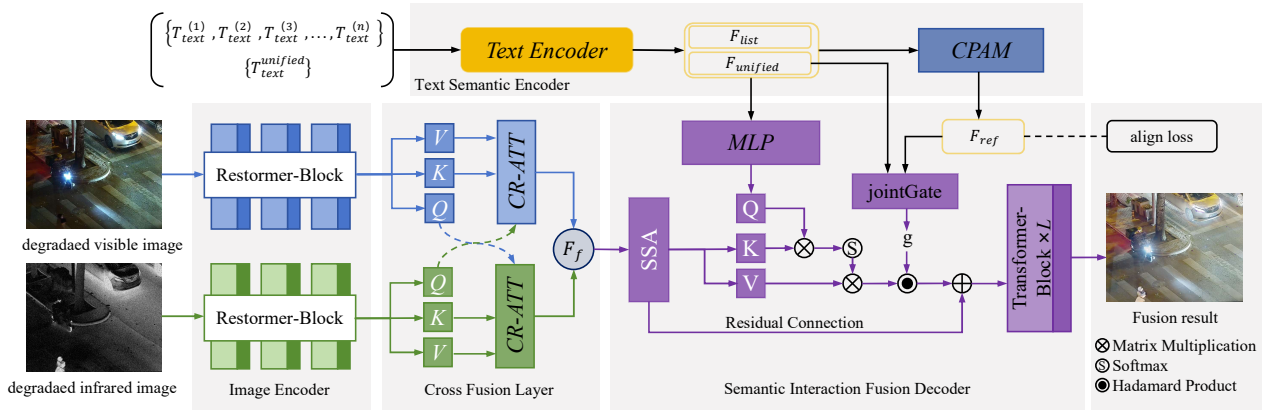

Fig. 1. The overall framework of our unified text-guided degradation-aware image fusion framework.

where $\mathcal{F}_{ir}$ and $\mathcal{F}_{vi}$ represent linear projection functions that map F_{ir} and F_{vi} to their corresponding Q, K, and V representations. Next, the queries Q from one modality are swapped to facilitate spatial interaction.

$$F_f^{ir} = \text{softmax}\left(\frac{Q_{vi}K_{ir}}{\sqrt{d_k}}\right)V_{ir}, \quad F_f^{vi} = \text{softmax}\left(\frac{Q_{ir}K_{vi}}{\sqrt{d_k}}\right)V_{vi}, \tag{5}$$

where d_k is the scaling factor. Finally, the cross-attended features are concatenated and fused via a convolution.

$$F_f = \text{Conv}(\text{Concat}(F_f^{ir}, F_f^{vi})). \tag{6}$$

Semantic Interaction Fusion Decoder. The fused features F_f are refined by spatial self-attention and further guided by text semantics F_{text} through the Semantic Guidance Module (SGM). The Transformer-based decoder $\mathcal{D}$ then reconstructs the fused image. Both encoder and decoder operate at multiple levels with downsampling and upsampling to capture rich details.

C. Text-Guided Module

The text-guided module couples semantic information from text with the fusion process, enhancing degradation-aware representation. Text serves as a high-level degradation condition to uniformly model both single and compound degradation types, and enhances the model's robustness and generalization in complex degradation scenarios by modulating cross-modal fusion features. Through the Cross-Prompt Attention Aggregation Module (CPAM), degradation-specific texts explicitly influence the feature modulation process via attention weights.

Textual Semantic Encoder. Given a text prompt T_{text}, we map it into an embedding using the frozen CLIP [17] text encoder E_t to ensure semantic consistency:

$$F_{text} = E_t(T_{text}), \tag{7}$$

where F_{text} denotes the textual semantic feature. Prompts with similar semantics yield embeddings that are close in Euclidean space.

Cross-Prompt Attention Aggregation Module(CPAM). To integrate multiple degradation prompts, we aggregate N sub-task prompts $F_{list} \in \mathbb{R}^{B\times N\times d}$ into a unified prompt $F_{unified} \in \mathbb{R}^{B\times d}$ via cross-attention. We use the unified prompt as the query (Q) and the sub-task prompts as keys (K) and values (V) for a multi-head attention layer. This enables the unified prompt to selectively attend to the most semantically relevant sub-task prompts, achieving effective multi-prompt integration:

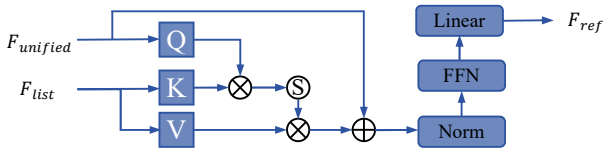

Fig. 2. The structure of Cross-Prompt Attention Aggregation Module(CPAM).

$$\begin{aligned} Q &= W_q F_{unified}, \quad K = W_k F_{list}, \quad V = W_v F_{list}, \\ attn &= Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V, \end{aligned} \tag{8}$$

where $attn$ denotes the output features of the cross-attention. We then apply residual connections and normalization, followed by a feed-forward network (FFN) and a linear mapping layer to obtain the reference feature F_{ref}:

$$F_{ref} = \text{Linear}(FFN(LN(F_{unified} + attn))). \tag{9}$$

F_{ref} supervises $F_{unified}$ during training, aligning multi-degradation tasks and providing the basis for gating.

Semantic Guidance Module (SGM). F_{text} extracted by the CLIP text encoder is first mapped into the visual space through an MLP $\phi(\cdot)$. Meanwhile, F_f from SSA is flattened and projected to the same dimension. By treating F_{text} as query and F_f as key/value, cross-modal attention is applied. This enables the text to selectively emphasize semantically relevant regions in the visual feature space, thereby generating enhanced features guided by textual semantics.

JointGate Module. To adaptively control the strength of text-guided visual enhancement, we design a JointGate Module. Based on $F_{unified}$ and F_{ref}, the module generates a dynamic gating vector $g \in [0,1]^{B\times d}$ that matches the

dimension of the visual features. The process is formulated as:

$$\begin{aligned} attn\_score &= \text{softmax}(\text{Linear}(\text{concat}(F_{unified}, F_{ref}))), \\ g &= \text{Sigmoid}(\text{Linear}(attn\_score \odot \text{concat}(F_{unified}, F_{ref}))) \end{aligned} \tag{10}$$

g balances the contributions of text and visual features, enabling adaptive semantic control under different degradations.

D. Loss Function for Two-Stage Training

To ensure stability and adaptability across multiple degradations, we adopt a two-stage training strategy with different loss designs for each stage.

Stage 1: Single-Degradation Training. Using infrared–visible image pairs under a single type of degradation, together with the corresponding textual prompts, the model learns fundamental degradation-aware enhancement and fusion capabilities. The Stage 1 loss $\mathcal{L}_{fusion}$ combines four terms:

The **intensity loss** $\mathcal{L}_{int}$ enhances overall brightness to highlight salient targets:

$$\mathcal{L}_{int} = \frac{1}{HW}\|I_f - \max(I_{ir}^g, I_{vi}^g)\|_1, \tag{11}$$

The **structural similarity loss** $\mathcal{L}_{ssim}$ preserves structural consistency:

$$\mathcal{L}_{ssim} = 2 - (\text{SSIM}(I_f, I_{ir}^g) + \text{SSIM}(I_f, I_{vi}^g)). \tag{12}$$

The **maximum gradient loss** $\mathcal{L}_{grad}$ enforces the preservation of critical edge information from both source images, leading to sharper textures in the fused output:

$$\mathcal{L}_{grad} = \frac{1}{HW}\|\nabla I_f - \max(\nabla I_{ir}^g, \nabla I_{vi}^g)\|_1, \tag{13}$$

The **color consistency loss** $\mathcal{L}_{color}$ keeps color fidelity with the visible image in YCbCr space:

$$\mathcal{L}_{color} = \frac{1}{HW}\|\mathcal{F}_{CbCr}(I_f) - \mathcal{F}_{CbCr}(I_{vi}^g)\|_1, \tag{14}$$

The final Stage 1 loss is a weighted sum:

$$\begin{aligned} \mathcal{L}_{fusion} = \alpha_{int}(t) \cdot \mathcal{L}_{int} + \alpha_{ssim}(t) \cdot \mathcal{L}_{ssim}(t) \\ + \alpha_{grad}(t) \cdot \mathcal{L}_{grad} + \alpha_{color}(t) \cdot \mathcal{L}_{color}, \end{aligned} \tag{15}$$

where $\alpha_*(t)$ are task-dependent weights modulated by the prompt t.

Stage 2: Multi-Degradation Alignment Training. For images with compound degradations, the reference feature F_{ref} is provided alongside the single-degradation prompts from Stage 1. Through CPAM, semantic alignment and knowledge sharing across degradation tasks are encouraged. To enforce this, we introduce a **cosine similarity alignment loss** $\mathcal{L}_{align}$:

$$\mathcal{L}_{align} = 1 - \frac{F_{unified} \cdot F_{ref}}{\|F_{unified}\|\|F_{ref}\|}, \tag{16}$$

The total Stage 2 loss is then defined as:

$$\mathcal{L}_{II} = \mathcal{L}_{fusion}^{unified} + \alpha \cdot \mathcal{L}_{align}, \tag{17}$$

where α is a balancing hyperparameter. $\mathcal{L}_{II}$ combines fusion and semantic alignment, enabling robust recovery under diverse degradations.

IV. Experiment

A. Implementation and Dataset

Implementation. For training, images are cropped into 96×96 patches with a batch size of 16. Stage 1 runs for 100 epochs on single-degradation pairs, and Stage 2 for 80 epochs on compound-degradation pairs, guided by corresponding textual prompts. We use AdamW with a learning rate of 1×10^{-4}, and all experiments are conducted on a Tesla V100 GPU using PyTorch. We use standardized degradation prompts. For example, for low-light scenarios: "This is the infrared–visible image fusion task. Visible images have low-light degradation." We then employ GPT-5 to generate hundreds of semantically consistent instructions for training, from which one is randomly selected at each iteration. These prompts do not need to precisely match the image content; semantic consistency is sufficient.

Dataset. We construct a dataset that includes four types of single degradations(low light(LL) and overexposure(OE) in visible images, and low contrast(LC) and random noise(RN) in infrared images) and four types of compound degradations(LC + LL, RN + LL, LC + OE, RN + OE). Our dataset is constructed by synthesizing typical degradations on existing high-quality infrared–visible fusion datasets (LLVIP [23], RoadScene [24], and M3FD [13]), including visible low-light, infrared noise, and infrared low-contrast conditions. We follow the mainstream paradigm of degradation-aware restoration and fusion datasets, as adopted in existing studies such as the EMS dataset proposed in Text-IF [3]. Visible low-light is simulated using a local illumination transmission-based degradation model. Infrared noise is generated by combining Gaussian, Poisson, and stripe noise with controllable parameters. Infrared low contrast is produced by compressing the effective dynamic range of the original histogram. In total, we obtain 3,880 training pairs and 1,488 testing pairs, each containing both clean and degraded infrared and visible images.

B. Fusion Performance Comparison

We compare our method with seven state-of-the-art (SOTA) fusion models: SwinFusion [9], YDTR [25], EMMA [26], Text-IF [3], SeAFusion [14], DCEvo [27], and DAFusion [28], where Text-IF and DAFusion jointly perform degradation restoration and fusion.

Without pre-enhancement. Table I reports results on standard fusion tasks. The consistently higher values of SSIM and VIF demonstrate that our method achieves superior perceptual fidelity. The higher values of EN and PSNR demonstrate that our results contain richer details and less distortion, leading to clearer and more reliable fused images.

With pre-enhancement. Figure III-C illustrates performance under both single degradations and compound degradations. For fair comparison, baseline fusion models are combined with advanced restoration networks—URetinex [29] for low-light image enhancement, NAFNet [30] for denoising, and AirNet [31] for contrast enhancement. Our method more effectively suppresses infrared noise on human subjects. In

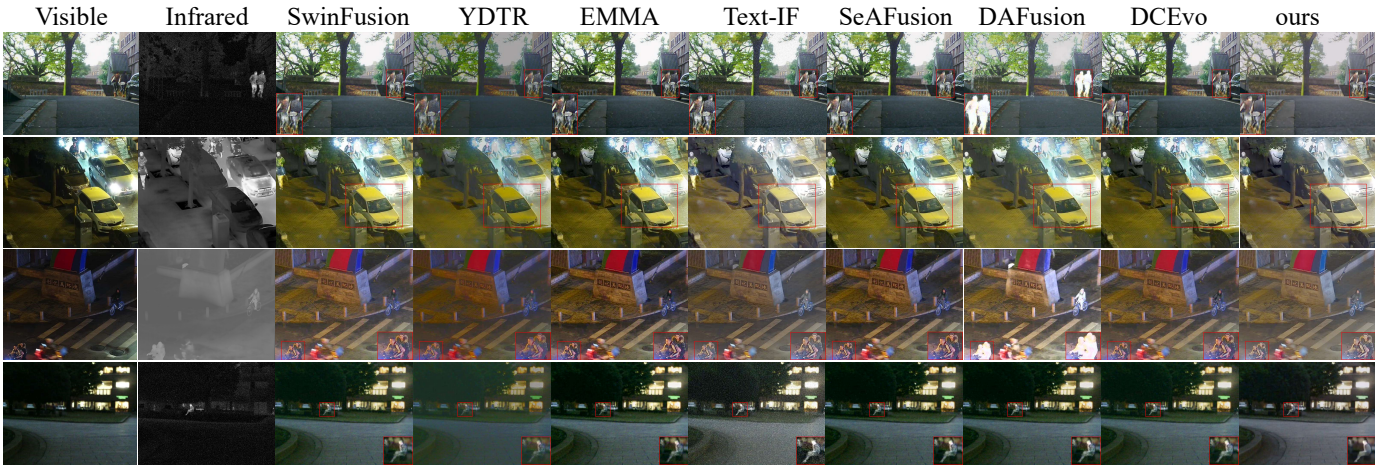

Fig. 3. Comparison of our method and the combination of existing image restoration and fusion methods on degraded source images. Degradations from top to bottom: RN, LL, LC + LL, and RN + LL. In RN and RN + LL degradation, our method more effectively suppresses infrared noise on human subjects. In LL degradation, it achieves higher fidelity, with vehicles rendered in correct colors.

TABLE I
QUANTITATIVE COMPARISON RESULTS ON TYPICAL FUSION DATASETS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND PURPLE.

Methods	TNO				RoadScene				VIFB			
	EN	SSIM	PSNR	VIF	EN	SSIM	PSNR	VIF	EN	SSIM	PSNR	VIF
SwinFusion	6.9	1.37	13.39	0.76	7.07	1.41	13.23	0.67	6.93	1.39	13.67	0.72
YDTR	6.42	1.39	14.32	0.6	6.8	1.41	14.7	0.59	6.63	1.4	14.94	0.58
EMMA	7.16	1.31	14.04	0.7	7.41	1.31	13.78	0.66	7.15	1.38	14.31	0.7
Text-IF	7.15	1.35	14.61	0.76	7.36	1.35	14.61	0.76	7.21	1.36	13.98	0.8
SeAFusion	7.1	1.32	13.74	0.69	7.44	1.3	13.69	0.68	7.03	1.36	13.74	0.66
DCEvo	6.95	1.37	13.5	0.83	7.16	1.38	13.51	0.75	6.97	1.38	13.68	0.74
DAFusion	7.43	1.17	12.31	0.72	7.49	1.31	14.22	0.67	7.42	1.25	12.66	0.84
ours	7.3	1.39	14.62	0.78	7.48	1.47	14.7	0.78	7.27	1.43	14.07	0.87

TABLE II
QUANTITATIVE COMPARISON RESULTS IN DIFFERENT DEGRADED SCENARIOS WITH PRE-ENHANCEMENT.

Methods	IR(RN)			VI(LL)			VI(LL) and IR(RN)			|VI(LL) and IR(LC)		
	EN	PSNR	CLIPIQA	EN	PSNR	CLIPIQA	EN	PSNR	CLIPIQA	EN	PSNR	CLIPIQA
SwinFusion	6.94	14.21	0.177	6.96	13.39	0.174	6.94	13.05	0.175	6.87	14.28	0.175
YDTR	6.22	14.03	0.153	6.25	13.54	0.172	6.22	12.91	0.154	6.04	13.33	0.162
EMMA	7.04	14.63	0.165	7.06	13.82	0.177	7.04	13.74	0.16	7.01	14.35	0.177
Text-IF	7.21	14.15	0.188	7.09	11.63	0.19	7.32	12.28	0.2	6.8	14.56	0.183
SeAFusion	7.03	14.59	0.171	7.05	13.73	0.182	7.03	13.64	0.173	7.0	14.58	0.182
DCEvo	6.95	14.23	0.185	6.99	13.51	0.182	6.95	13.07	0.184	6.91	14.38	0.184
DAFusion	7.33	10.23	0.173	7.34	9.4	0.174	7.33	9.35	0.177	7.4	10.34	0.185
ours	7.43	14.6	0.186	7.3	13.77	0.2	7.43	13.99	0.197	7.41	14.57	0.188

LL degradation, it achieves higher fidelity, with vehicles rendered in correct colors. As shown in Table II, our method demonstrates significant advantages in compound degradation scenarios. Under the VI(LL)+IR(RN) condition, Ours achieves the highest scores in both Information Entropy (EN, 7.43) and Peak Signal-to-Noise Ratio (PSNR, 13.99), while also attaining a competitive CLIPIQA score (0.197). This indicates that our approach enhances the visual naturalness and semantic consistency of the fused results while preserving rich detail and structural fidelity.

C. Ablation Study

We design four ablation settings to assess the contribution of each component: (I) **w/o** two-stage training (trained only on compound degradations); (II) **w/o** JointGate Module; (III) **w/o** Cross-Prompt Attention Module (CPAM), using only $unified\_text$ in Stage 2; (IV) **w/o** $\mathcal{L}_{align}$.

The quantitative results reported in Table 4 demonstrate that the complete model (Ours) outperforms all ablation variants in both mixed degradation scenarios—visible low-light with infrared random noise and visible low-light with infrared low-contrast—on the metrics of Entropy (EN) and Peak Signal-to-Noise Ratio (PSNR), confirming the necessity of the overall architectural design. The most significant performance drop occurs when removing the two-stage training strategy (Config I), notably reducing PSNR from 13.99 to 12.5 in the VI(LL)+IR(RN) setting, which indicates the crucial role of progressive learning in establishing a solid foundation before tackling complex compound degradations. The absence of the JointGate module (Config II) leads to a marked decline, especially in the perception-aware CLIPIQA metric, demonstrating its essential function in dynamically gating and adaptively fusing cross-modal features to integrate critical information from both modalities. While replacing the Cross-Prompt Attention Module (CPAM) with a single unified text prompt (Config III) yields better performance than Configs I and II, it remains substantially lower than the full model, highlighting the value of fine-grained, degradation-specific feature modulation enabled by CPAM. Finally, removing the text-image alignment loss$\mathcal{L}_{align}$ (Config IV) results in the second-best performance among ablated configurations, yet a clear gap remains compared to the full model, suggesting that $\mathcal{L}_{align}$ acts as a high-level regularizer that enhances performance by enforcing semantic consistency between the fused output and the textual condition.

TABLE III
QUANTITATIVE RESULTS OF THE ABLATION STUDIES.

Configs	VI(LL) and IR(RN)			VI(LL) and IR(LC)		
	EN	PSNR	CLIPIQA	EN	PSNR	CLIPIQA
I	6.96	12.5	0.191	6.85	13.45	0.18
II	6.85	13.45	0.181	6.81	12.99	0.175
III	6.9	13.5	0.186	6.98	13.6	0.182
IV	7.2	13.6	0.19	7.3	13.8	0.186
ours	7.43	13.99	0.197	7.41	14.57	0.188

V. CONCLUSION

We proposed a text-guided degradation-aware framework for infrared–visible image fusion. With Cross-Prompt Attention, a JointGate Module, and text-guided attention, our method achieves adaptive and robust fusion under diverse degradations. We further built a dataset covering multiple types of degradation, which provides a valuable benchmark and foundation for advancing academic research in multi-degradation fusion and cross-modal learning. In future work, we will extend this framework to broader degradation conditions and cross-modal vision tasks.

REFERENCES

[1] Xuan Li, Zhenghua Huang, Lei Ma, Yuhang Xu, Li Cheng, and Zhi Yang, "Reliable metrics-based linear regression model for multilevel privacy measurement of face instances," *IET Image Processing*, vol. 16, no. 7, pp. 1935–1948, 2022.

[2] Xuan Li, Yuhang Xu, Zhenghua Huang, Lei Ma, and Zhi Yang, "The extraction of pixel-wise visual multi-cues for ahp-based privacy measurement," *Optik*, vol. 249, pp. 168238, 2022.

[3] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma, "Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *CVPR*, 2024, pp. 27026–27035.

[4] Jiayi Ma, Linfeng Tang, Meilong Xu, Hao Zhang, and Guobao Xiao, "Stdfusionnet: An infrared and visible image fusion network based on salient target detection," *IEEE TIM*, vol. 70, pp. 1–13, 2021.

[5] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu, "Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *CVPR*, 2023, pp. 13955–13965.

[6] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler, "Lrrnet: A novel representation learning guided fusion network for infrared and visible images," *IEEE TPAMI*, 2023.

[7] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *CVPR*, 2023, pp. 5906–5916.

[8] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang, "Fusiongan: A generative adversarial network for infrared and visible image fusion," *InfFus*, vol. 48, pp. 11–26, 2019.

[9] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA JAS*, vol. 9, no. 7, pp. 1200–1217, 2022.

[10] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool, "Ddfm: denoising diffusion model for multi-modality image fusion," in *ICCV*, 2023.

[11] Han Xu, Jiteng Yuan, and Jiayi Ma, "MURF: mutually reinforcing multi-modal image registration and fusion," *IEEE TPAMI*, vol. 45, no. 10, pp. 12148–12166, 2023.

[12] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *IJCAI*, 2022, pp. 3508–3515.

[13] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *CVPR*, 2022, pp. 5802–5811.

[14] Linfeng Tang, Jiteng Yuan, and Jiayi Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *InfFus*, vol. 82, pp. 28–42, 2022.

[15] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma, "Divfusion: Darkness-free infrared and visible image fusion," *Information Fusion*, vol. 91, pp. 477–493, 2023.

[16] Zhu Liu, Jinyuan Liu, Benzhuang Zhang, Long Ma, Xin Fan, and Risheng Liu, "Paif: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation," *arXiv preprint arXiv:2308.03979*, 2023.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[18] Vaishnav Potlapalli, SyedWaqas Zamir, Salman Khan, and FahadShahbaz Khan, "Promptir: Prompting for all-in-one blind image restoration," Jun 2023.

[19] Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu, "Autodir: Automatic all-in-one image restoration with latent diffusion," Oct 2023.

[20] MarcosV. Conde, Gregor Geigle, and Radu Timofte, "Instructir: High-quality image restoration following human instructions," Feb 2024.

[21] Chenyang Qi, Zhengzhong Tu, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Qifeng Chen, and Hossein Talebi, "Spire: Semantic prompt-driven image restoration," Jul 2024.

[22] Ziwei Luo, FredrikK Gustafsson, Zheng Zhao, Jens Sjölund, and ThomasB Schön, "Controlling vision-language models for universal image restoration," .

[23] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *ICCV*, 2021, pp. 3496–3504.

[24] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo, "Fusiondn: A unified densely connected network for image fusion," in *AAAI*, 2020, vol. 34, pp. 12484–12491.

[25] Wei Tang, Fazhi He, and Yu Liu, "Ydtr: Infrared and visible image fusion via y-shape dynamic transformer," *IEEE TMM*, 2022.

[26] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool, "Equivariant multi-modality image fusion," in *CVPR*, 2024, pp. 25912–25921.

[27] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan, "Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2226–2235.

[28] Xue Wang, Zheng Guan, Wenhua Qian, Jinde Cao, Runzhuo Ma, and Cong Bi, "A degradation-aware guided fusion network for infrared and visible image," *Information Fusion*, vol. 118, pp. 102931, 2025.

[29] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang, "Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5901–5910.

[30] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun, "Simple baselines for image restoration," in *European conference on computer vision*. Springer, 2022, pp. 17–33.

[31] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng, "All-in-one image restoration for unknown corruption," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17452–17462.