Robust Fusion Controller: Degradation-Aware Image Fusion with Fine-Grained Language Instructions

Hao Zhang[1,2*], **Yanping Zha**[1*], **Qingwei Zhuang**[3], **Zhenfeng Shao**[3], **Jiayi Ma**[1†]

[1]Electronic Information School, Wuhan University, China
[2]Suzhou Institute of Wuhan University, China
[3]State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, China
{zhpersonalbox, yanpingcha66, jyma2010}@gmail.com, {zhuangqingwei, shaozhenfeng}@whu.edu.cn

Abstract

Current image fusion methods struggle to adapt to real-world environments encompassing diverse degradations with spatially varying characteristics. To address this challenge, we propose a robust fusion controller (RFC) capable of achieving degradation-aware image fusion through fine-grained language instructions, ensuring its reliable application in adverse environments. Specifically, RFC first parses language instructions to innovatively derive the functional condition and the spatial condition, where the former specifies the degradation type to remove, while the latter defines its spatial coverage. Then, a composite control priori is generated through a multi-condition coupling network, achieving a seamless transition from abstract language instructions to latent control variables. Subsequently, we design a hybrid attention-based fusion network to aggregate multi-modal information, in which the obtained composite control priori is deeply embedded to linearly modulate the intermediate fused features. To ensure the alignment between language instructions and control outcomes, we introduce a novel language-feature alignment loss, which constrains the consistency between feature-level gains and the composite control priori. Extensive experiments on publicly available datasets demonstrate that our RFC is robust against various composite degradations, particularly in highly challenging flare scenarios.

Code — https://github.com/HaoZhang1018/RFC

Introduction

Due to the limitation of the imaging principle, single-modal images can only capture partial scene attributes, failing to support comprehensive perception. In this context, image fusion technology emerges (Singh et al. 2023; Huang et al. 2024; Liu et al. 2024a,c), aiming to integrate complementary information from multi-modal images to provide a comprehensive representation of the imaging scene. Thanks to this excellent representational capability, image fusion has become a core component of numerous intelligent perception applications, effectively enhancing the accuracy of military reconnaissance (Muller and Narayanan 2009), autonomous driving (Yadav et al. 2020), *etc*.

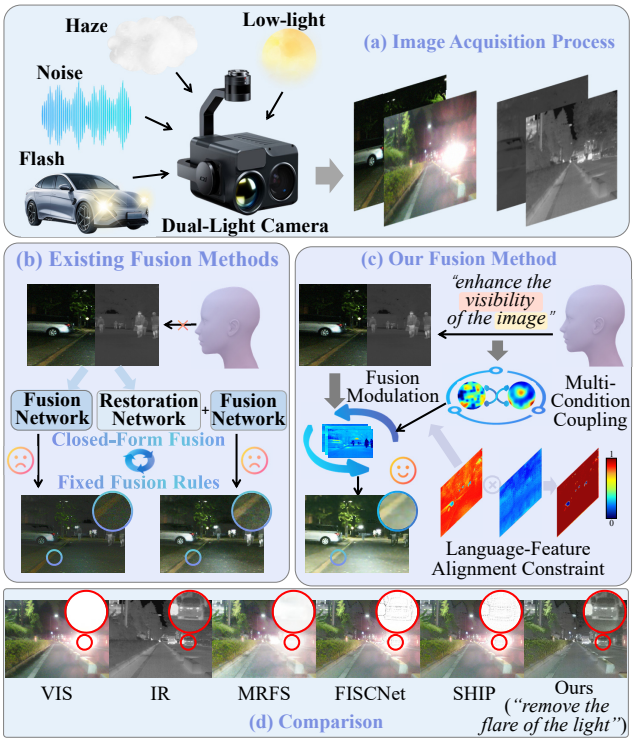

Figure 1: Comparisons between our RFC and existing fusion methods in the degraded scenario.

From the task definition, the application scenarios of image fusion typically involve environments where a single sensor is ineffective due to poor conditions. In the real world, such environments typically exhibit two key characteristics. On the one hand, degradations are pervasive (*e.g.*, overexposure, low-light, noise, haze, blur, and flare), with their types being diverse and severely compounded. On the other hand, these degradations exhibit spatially varying characteristics, potentially occurring both globally and locally. For instance, noise often appears in low-illumination regions, while flares typically accompany light sources. Therefore, equipping fusion models with the ability to overcome spatial-varying composite degradations is crucial for ensuring their reliable application in the real world.

*These authors contributed equally.
†Corresponding author

Unfortunately, existing fusion methods struggle to meet this requirement, fundamentally hindering the practical application of image fusion technology. More concretely, mainstream fusion methods (Li and Wu 2018; Zhang and Demiris 2023; Ma et al. 2019; Tang, Yuan, and Ma 2022; Liu et al. 2024d) that focus solely on enhancing information aggregation capabilities essentially do not eliminate degradations. Instead, the persistent presence of degradations leads to the erroneous discarding of valuable information, rendering image fusion more akin to a problem of "*information compression*". Differently, some of the latest methods (Zhang et al. 2024a; Chen et al. 2024; Zou and Yang 2023; Zhang et al. 2024c; Tang et al. 2023) cooperate to achieve degradation removal and information fusion, enabling the restoration of more information from low-quality source images. This perspective tends to turn image fusion into a problem of "*information mining*". However, these methods can only handle a single type of degradation and are ineffective against composite degradations, let alone those with spatial variability.

To address these challenges, we propose a robust fusion controller, termed RFC. It derives a degradation-aware image fusion framework with fine-grained language instructions, enabling adaptability to harsh environments with spatial-varying composite degradations. Firstly, RFC parses language instructions to obtain two complementary control conditions. 1) ***Functional condition***: enables the specification of the degradation type to be removed, supporting both single-type degradation removal and unified removal of composite degradations. 2) ***Spatial condition***: defines the regions to be enhanced, supporting both local and global enhancement. Secondly, functional and spatial conditions are processed through a multi-condition coupling network, to generate composite control priori. This process translates abstract language instructions into latent control variables, providing a high-quality interactive medium for dynamically modulating the fusion process. Thirdly, the composite control priori is embedded into a hybrid attention-based fusion network through the linear feature modulation strategy (Perez et al. 2018). While aggregating multi-modal information, it can precisely perceive and remove spatial-varying composite degradations. Finally, a novel language-feature alignment loss is introduced. By constraining the consistency between feature-level gains and the composite control priori, it can ensure that the controlled output aligns with the expectations of the language instructions. As presented in Fig. 1, our RFC significantly outperforms state-of-the-art methods in terms of harsh scenario characterization, particularly in challenging flare environments.

In summary, we make the following contributions:

- We propose a robust fusion controller, forming a degradation-aware image fusion framework with fine-grained language instructions. To our knowledge, this is the first attempt in the field of image fusion to eliminate spatial-varying composite degradations, enhancing the robustness of fusion models in harsh environments.
- We design a novel generative mechanism for composite control priori, which can translate abstract language instructions into latent control variables. This enables us to establish an open-ended paradigm for image fusion, facilitating fine-grained functional control over arbitrary regions in accordance with user-defined instructions.
- A language-feature alignment loss is introduced, which drives feature gains to maintain potential consistency with composite control priori, strongly ensuring the modulation rationality of our RFC.

Methodology

Our RFC leverages language instructions to guide fusion, ensuring high-quality multi-modal aggregation while accurately removing spatial-varying composite degradations. We first parse instructions into functional and spatial conditions, defining the desired operation and target regions. These are then coupled into a composite control priori, modulating hybrid attention fusion modules to achieve the desired results. The overall framework is shown in Fig. 2.

Language Instruction Parsing

Given the input language instruction ζ, which expresses a composite requirement. First, we split ζ to obtain language fragments ζ_f that describe the functions (*e.g.*, remove noise) and language fragments ζ_s that specify the spatial regions. Their semantic content differs significantly, so we introduce two strategies to parse them separately. Specifically, ζ_f essentially represents a requirement for visual appearance, so we leverage the visual-text alignment capability of the CLIP (Radford et al. 2021) model to parse it:

$$\alpha = E_T(\zeta_f), \tag{1}$$

where E_T is the text encoder from the pretrained CLIP, α indicates the obtained functional condition.

In contrast, ζ_s is more related to spatial localization, which cannot be handled by CLIP due to a lack of fine-grained parsing capability. Thus, we use a powerful spatial parsing model, CLIPSeg (Lüddecke and Ecker 2022), for analyzing language fragments ζ_s, to locate specific image regions based on language instructions: $\{S_{vis}, S_{ir}\} = \text{CSeg}(\{I_{vis}, I_{ir}\}|\zeta_s)$, where CSeg indicates the function of the pertained CLIPSeg, $\{I_{vis}, I_{ir}\}$ denotes the visible and infrared image pairs, and $\{S_{vis}, S_{ir}\}$ represent the output spatial response maps. Considering that CLIPSeg is trained only on the visible modality, the location confidence on multi-modal data may be reduced. Thus, we fine-tune the CLIPSeg by unfreezing the parameters in the last convolutional layers Φ_c of its decoder. The fine-tuned spatial response maps are generated by $\{S^{'}_{vis}, S^{'}_{ir}\} = \text{CSeg}_{\Phi^{'}_c}(\{I_{vis}, I_{ir}\}|\zeta_s)$, where $\text{CSeg}_{\Phi^{'}_c}$ is the fine-tuned CLIPSeg. To combine the cross-modal priori knowledge before and after fine-tuning, we perform spatial response mixing, obtaining a comprehensive spatial condition β:

$$\beta = S_{vis} \oplus S_{ir} \oplus S^{'}_{vis} \oplus S^{'}_{ir}, \tag{2}$$

where $\oplus$ is the concatenation operation. Now, through parsing the input language instruction, we obtain the functional condition α and the spatial condition β.

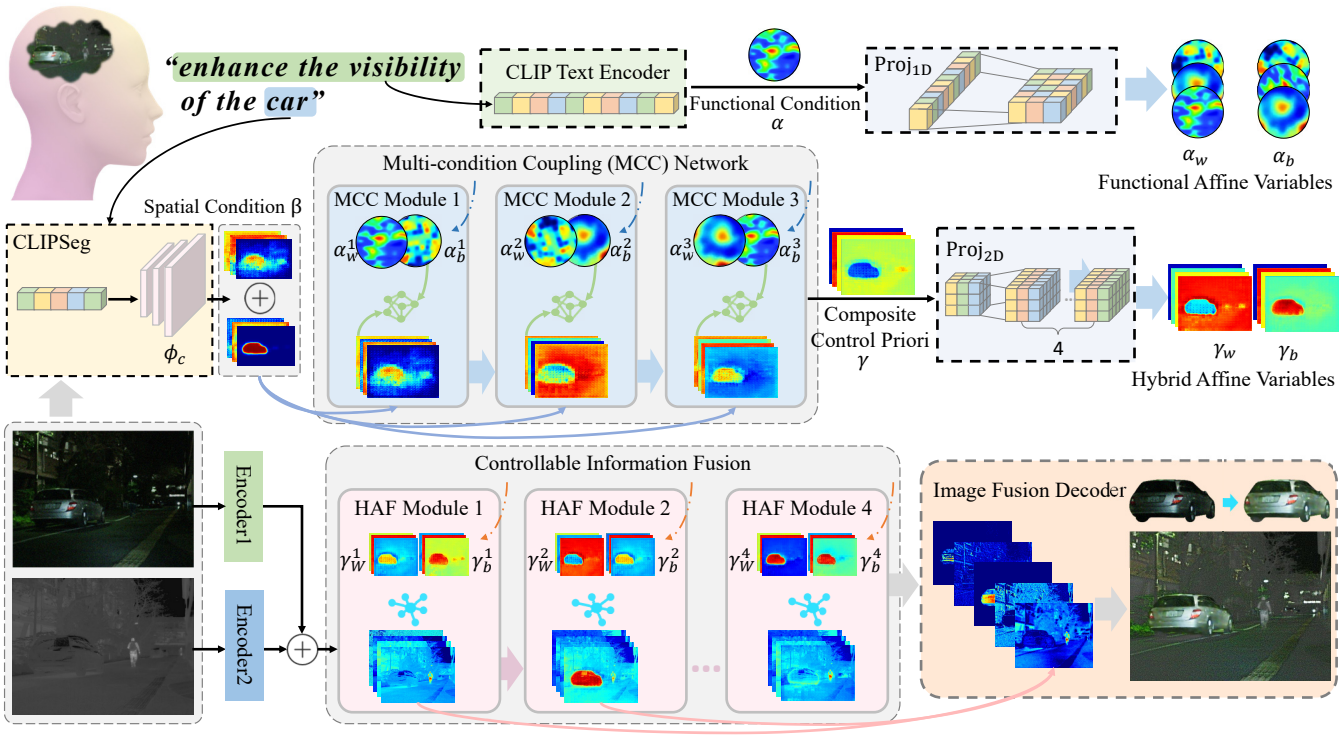

Figure 2: The overall framework of our proposed RFC.

Composite Control Priori Generation

Effective control of the fusion process necessitates a variable that captures both functional and spatial conditions. Inspired by FiLM (Perez et al. 2018), we design multi-condition coupling (MCC) modules to combine α and β following the idea of feature-wise affine transformation, as shown in Fig. 3.

We first use two 1D convolution layers to generate functional affine variables: $\{\alpha_w^i, \alpha_b^i\} = \text{Proj}_{1D}^i(\alpha)$, where α_w^i indicates the functional weight, α_b^i denotes the functional bias, and i is the index of the multi-condition coupling module. Then, we perform an affine transformation (AT) for a functional compound and combine it with the spatial condition: $\text{AT}(\cdot|\alpha_w^i, \alpha_b^i) \oplus \beta$. Such an operation ensures that the functional and spatial conditions are fully coupled, serving as the core component of the MCC module. The function of the MCC module can be represented as:

$$F_{out}^i = \text{MCC}(F_{out}^{i-1}, \text{AT}(\cdot|\alpha_w^i, \alpha_b^i) \oplus \beta), \tag{3}$$

where F_{out}^{i-1} is the composite control priori output from the $i-1$-th MCC module, and when $i = 1$, $F_{out}^{i-1} = \beta$. Totally, 3 MCC modules are used, and the output F_{out}^3 of the last module is regarded as the composite control priori γ. It then undergoes 2D convolution to produce final hybrid affine variables $\{\gamma_w^k, \gamma_b^k\} = \text{Proj}_{2D}^k(\gamma)$, which can be considered to fully integrate both functional and spatial conditions.

Controllable Information Fusion

Next, the task at hand is to enable high-quality multi-modal feature fusion and seamlessly embed the generated composite control priori into the fusion process. We develop hybrid attention fusion (HAF) modules based on CBAM (Woo et al. 2018) to achieve this goal, as illustrated in Fig. 4. First, we employ the channel attention mechanism to blend infrared and visible features. Formally, we use pooling operations to squeeze the input feature F, obtaining the maximum and average responses along the spatial dimensions, respectively. These responses are processed by a multi-layer perceptron (MLP), combined through summation, and subjected to a nonlinear activation to produce the final attention map, which is used to enhance the aggregated feature F. This process can be represented as:

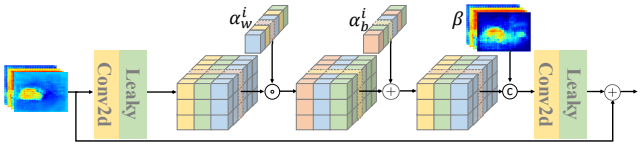

Figure 3: The structure of multi-condition coupling module.

$$F_c^k = \sigma(\kappa(P_M(F^{k-1})) + \kappa(P_A(F^{k-1}))) \otimes F^{k-1}, \tag{4}$$

where F^{k-1} is the feature output from the $k-1$-th HAF module, and when $k = 1$, $F^{k-1} = F_{ir} \oplus F_{vis}$. P_M and P_A denote the maximum and average pooling, κ indicates the MLP function, and σ indicates the Sigmoid function. Building on it, the spatial attention mechanism is utilized to reinforce the spatial representation of fused features F_c^k. Concretely, pooling operations are applied to F_c^k to extract the maximum and average responses in the channel dimen-

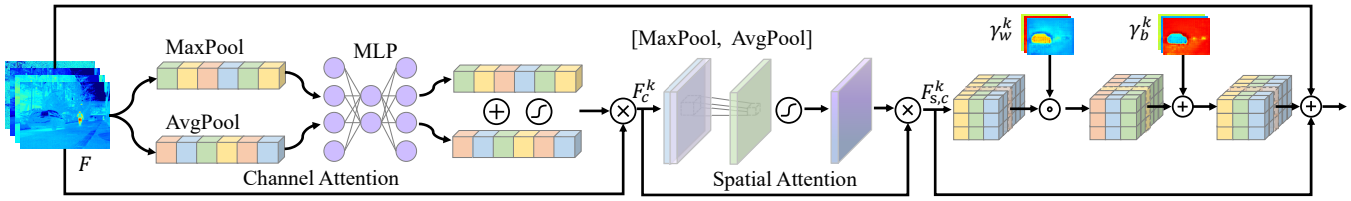

Figure 4: The structure of hybrid attention fusion module.

sion, which are subsequently concatenated, projected, and activated to produce the spatial attention map. The spatial attention-based reinforcement process can be defined as:

$$F_{c,s}^k = \sigma(\kappa(P_A(F_c^k) \oplus P_M(F_c^k))) \otimes F_c^k, \tag{5}$$

where $F_{c,s}^k$ is the fused feature that has been attention-enhanced across both the spatial and channel dimensions. For embedding the composite control priori, we still follow the idea of feature-wise affine transformation. Specifically, we use the affine variables $\{\gamma_w^k, \gamma_b^k\}$ to process $F_{c,s}^k$:

$$F_{control}^k = \mathrm{AT}(F_{c,s}^k | \gamma_w^k, \gamma_b^k) + F^{k-1}. \tag{6}$$

Finally, an UNet-like (Ronneberger, Fischer, and Brox 2015) decoder D_U with skip connections is employed to reconstruct the fused image: $I_f = D_U(F_{control}^1, F_{control}^2, F_{control}^3, F_{control}^4)$.

Optimization Regularization

The above designs offer the architectural support for robust image fusion with fine-grained language instructions. To ensure their effective operation, we formulate optimization regularization, comprising a degradation-aware reconstruction loss and a language-feature alignment loss.

Degradation-Aware Reconstruction Loss. This regularization term aims to drive the targeted removal of degradations, enhancing perceptual fidelity. The data used to construct this loss is multi-modal clean-degraded image pairs $\{I_{vis}, I_{ir}, I_{vis}^{'}, I_{ir}^{'}\}$, where I_{vis} and I_{ir} are degraded images, and $I_{vis}^{'}$ and $I_{ir}^{'}$ are corresponding clean ones.

Based on the input language instruction ζ, we identify two conditions: the degradation type Ω (*e.g*, low-light, flare, haze, noise, blur, and their composites) and the target region Λ (can be either a local region or the entire image). Firstly, according to Ω, we retrieve $\{I_{vis}^{\Omega}, I_{ir}^{\Omega}\}$ that includes this specific degradation (or a compound of multiple types of degradation) from the dataset. Secondly, in conjunction with Λ, we simulate pseudo multi-modal references:

$$\{\hat{I}_{vis}, \hat{I}_{ir}\} = \{I_{vis}^{\Omega}, I_{ir}^{\Omega}\}_{\overline{\Lambda}} + \{I_{vis}^{'}, I_{ir}^{'}\}_{\Lambda}, \tag{7}$$

where $\overline{\Lambda}$ indicate regions that are not specified by language instruction ζ. With the pseudo multi-modal references in place, we construct the degradation-aware reconstruction loss to constrain the final fused image I_f from three aspects: contrast, structure, and color: $\mathcal{L}_{rec} = \mathcal{L}_{con} + \mathcal{L}_{str} + \mathcal{L}_{cor}$. The corresponding loss terms are defined as:

$$\mathcal{L}_{con} = \sum \alpha_{\{\Lambda,\overline{\Lambda}\}} \| I_f^y - \max(\{\hat{I}_{vis}^y, \hat{I}_{ir}\}) \|_{\{\Lambda,\overline{\Lambda}\}}, \tag{8}$$

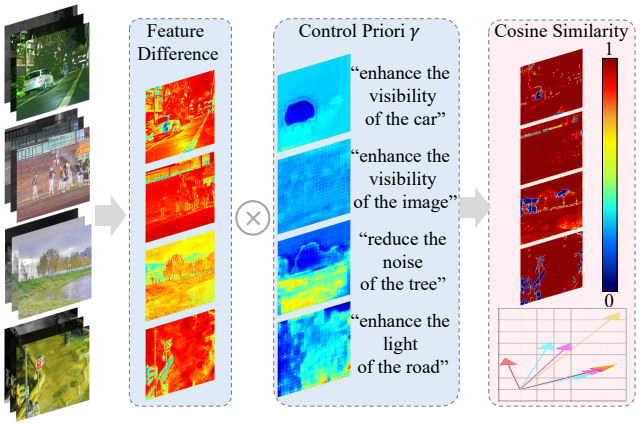

Figure 5: Schematic diagram of the alignment mechanism between feature-level gains and language instructions.

$$\mathcal{L}_{str} = \sum \alpha_{\{\Lambda,\overline{\Lambda}\}} \| \nabla I_f^y - \max(\{\nabla \hat{I}_{vis}^y, \nabla \hat{I}_{ir}\}) \|_{\{\Lambda,\overline{\Lambda}\}}, \tag{9}$$

$$\mathcal{L}_{cor} = \sum \alpha_{\{\Lambda,\overline{\Lambda}\}} \| I_f^{cbcr} - \hat{I}_{vis}^{cbcr} \|_{\{\Lambda,\overline{\Lambda}\}}, \tag{10}$$

where superscripts y and $cbcr$ denote the illumination and chrominance channels, respectively. We use dynamic weights $\alpha_{\{\Lambda,\overline{\Lambda}\}}$ to distinctively handle the distance calculations for the language-specified region Λ and other regions $\overline{\Lambda}$. The dynamic weights are defined as $\alpha_{\Lambda} = \Upsilon(I_f)/\Upsilon(\Lambda)$, in which Υ is an operator that calculates the number of pixels in specific regions. Such a mechanism can prevent the limitation of the target region from being overlooked during the optimization process when its size is too small.

Language-Feature Alignment Loss. Through the above loss, the dynamic responsiveness of the final fused image to language instructions can be effectively driven. However, the internal fusion process still lacks constraints, which potentially compromises the fusion model's sensitivity to language instructions. Thus, we introduce a novel language-feature alignment loss, primarily ensuring that the feature-level gains introduced by HAF modules remain consistent with the composite control priori. As shown in Fig. 5, we calculate the residual between the input of the first HAF module and the output of the final HAF module, representing the feature-level gains by the composite control priori: $\Delta F = F_{control}^0 - F_{control}^4 = F_{ir} \oplus F_{vis} - F_{control}^4$. Then, the language-feature alignment loss is defined as:

$$\mathcal{L}_{ali} = 1 - \frac{\langle \tau(\gamma), \tau(\Delta F) \rangle}{|\tau(\gamma)| \times |\tau(\Delta F)|}, \tag{11}$$

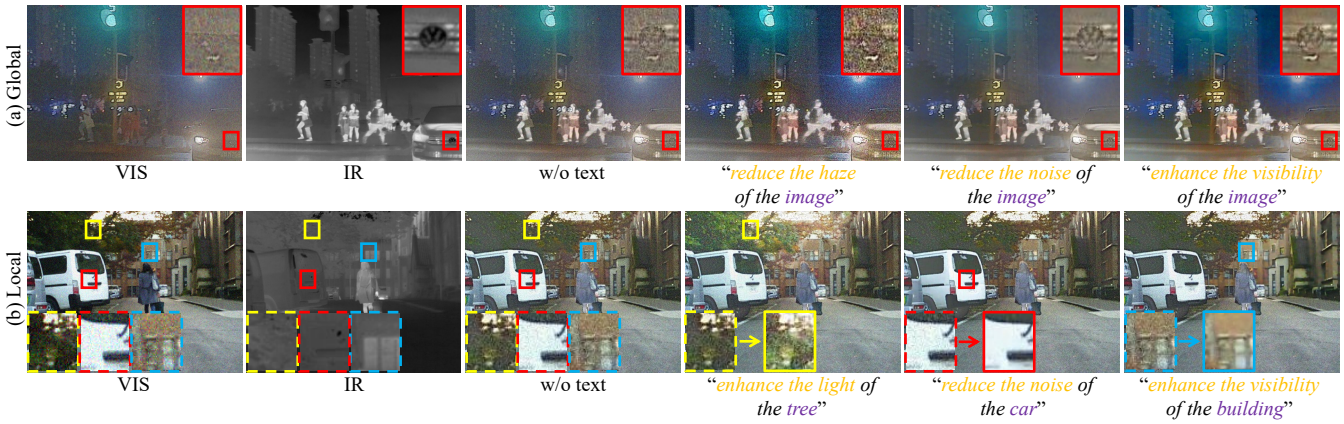

Figure 6: Demonstrating global and local degradation-aware fusion ability of our RFC.

where τ is the flattening operator, and $\langle \cdot \rangle$ is the vector dot product. This loss effectively ensures the rationality of the intermediate fusion processes under language instructions.

Experiments

Experimental Configurations

Datasets. We construct the required training and testing dataset based on MFNet (Ha et al. 2017), LLVIP (Jia et al. 2021), M3FD (Liu et al. 2022), FMB (Liu et al. 2023), and RoadScene (Xu et al. 2020) datasets. Specifically, we extend these existing datasets with simulated degradations (*e.g*, low light, flare, haze, noise, blur, and their composites). Our training set includes $14,654$ image-text pairs with annotations specifying degradation types and regions. Testing employs 700 multi-modal image pairs.

Implementation. We use the AdamW optimizer with an initial learning rate $2e^{-4}$ to update the parameters of all network modules. Meanwhile, the multi-scale training strategy is adopted to enhance our RFC's generalization performance across images of varying scales. All experiments are conducted on four NVIDIA Tesla P100 GPUs with 16 GB memory and one Intel(R) Xeon(R) Gold 5117 CPU.

Robust Fusion Controller Validation

First, we demonstrate our method's capability as a robust fusion controller that effectively removes degradation artifacts during the fusion process, both globally and locally.

Global Degradation-Aware Fusion. As shown in Fig. 6 (a), our RFC achieves global degradation removal under different language instructions. For instance, instructions like "*reduce the noise*" selectively suppress noise while preserving other characteristics, whereas "*enhance the visibility*" eliminates composite degradations, producing a completely clean output. This highlights RFC's capability to interpret language instructions for targeted enhancement precisely.

Local Degradation-Aware Fusion. Our RFC also supports local degradation removal, addressing practical needs such as enhancing key objects (*e.g.*, cars, pedestrians). As shown in Fig. 6 (b), when an instruction specifies both the degradation type and target region, RFC selectively restores the designated areas while maintaining contextual consistency.

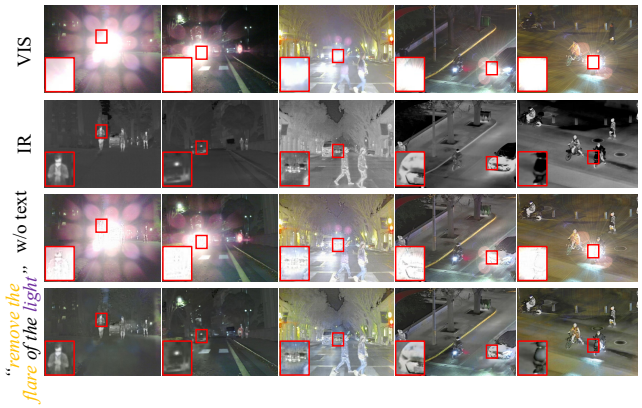

Figure 7: The flare removal function of our RFC.

Flare Removal. In nighttime driving scenarios, lens flare degrades image quality, impairing visibility. As a highlight, our RFC mitigates flare artifacts through language-driven modulation in Fig. 7, leveraging infrared cues to compensate for overexposed regions. This results in robustness improvements to produce perceptually superior fused results.

Comparison Under Composite Degradations

We first compared RFC with nine SOTA methods under composite-degradation scenarios, including MRFS (Zhang et al. 2024b), CDDFuse (Zhao et al. 2023a), DDFM (Zhao et al. 2023b), CAF (Liu et al. 2024b), SHIP (Zheng et al. 2024a), FISCNet (Zheng et al. 2024b), ReFusion (Bai et al. 2024), SDCFusion (Liu et al. 2024e), and Text-IF (Yi et al. 2024). For methods without degradation removal capabilities, we use an all-in-one enhancement method, InstructIR (Conde, Geigle, and Timofte 2024) for pre-processing. For Text-IF, we inform it of all the types of degradation present in the source images through textual input. Our RFC is tested with the default instruction: "enhance the visibility of the image". Visual results in Fig. 8 highlight RFC's advantages in handling composite degradations. For example, the competitors all fail to handle the noise that is introduced

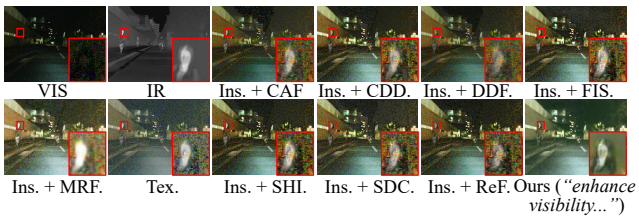

Figure 8: Qualitative results under composite degradations.

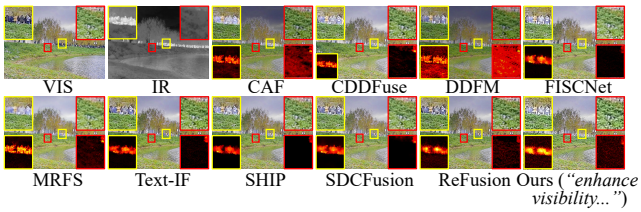

Figure 9: Qualitative results under no degradation. The highlighted area at the bottom represents the residual maps between fused results and the source visible image.

Degradation	DDF.	SHI.	CDD.	MRF.	CAF	ReF.	FIS.	SDC.	Tex.	Ours
Qabf	0.25	0.41	0.39	0.24	0.32	0.43	0.43	0.43	0.38	**0.45**
SSIM	0.27	0.27	0.29	0.25	0.28	0.29	0.27	0.28	0.23	**0.32**
MI	2.29	2.53	2.73	2.47	2.34	2.62	2.40	2.38	2.22	**2.79**
VIF	0.44	0.45	0.48	0.42	0.44	0.49	0.46	0.49	0.39	**0.50**
SCD	**1.35**	1.06	1.27	1.16	1.25	1.26	1.04	1.24	1.04	1.28

Table 1: Quantitative results under composite degradations.

No-Degra.	DDF.	SHI.	CDD.	MRF.	CAF	ReF.	FIS.	SDC.	Tex.	Ours
SD	37.18	46.54	49.16	46.83	39.65	48.19	48.63	47.04	**50.89**	49.90
AG	4.40	7.37	6.96	5.16	6.23	7.30	7.79	7.32	7.62	**9.78**
EN	6.99	7.24	7.28	7.24	7.03	7.30	7.31	7.29	7.35	**7.36**
SF	12.48	21.28	20.71	14.66	21.65	21.27	22.26	21.07	22.14	**26.62**

Table 2: Quantitative results under no degradation.

when enhancing illumination. In contrast, our RFC effectively removes composite degradations, while preserving the saliency of the pedestrian and the fine details in the background. Furthermore, we use five full-reference metrics to compute the correlation between the fused result and source images, as shown in Table 1. Our RFC outperforms other methods on most metrics, showing its ability to retain key scene information.

Comparison Under No Degradation

Beyond removing degradations, our RFC can further generate additional textures in degradation-free environments with the instruction "enhance the details of the image". As shown in Fig. 9, the residual maps between fused results and the source visible image reveal that RFC retains more high-frequency information in areas like trees and grass, while preserving richer thermal radiation in the human region. Since *RFC achieves information generation beyond source images in degradation-free scenarios, full-reference metrics in Table 1 are no longer applicable. Instead, we employ four no-reference metrics for objective evaluation*. In Table 2, our method achieves the best scores on most metrics.

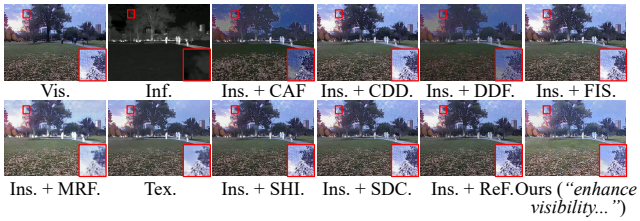

Figure 10: Qualitative results of generalization experiment.

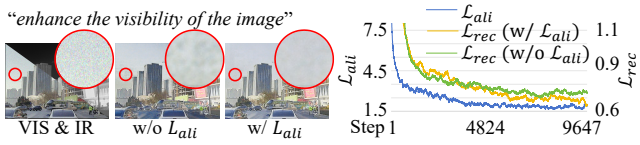

Figure 11: Ablation on language-feature alignment loss.

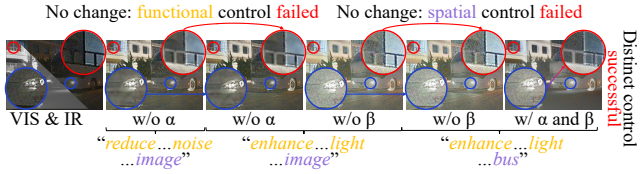

Figure 12: Ablation on functional and spatial conditions.

General.	DDF.	SHI.	CDD.	MRF.	CAF	ReF.	FIS.	SDC.	Tex.	Ours
Qabf	0.30	0.48	0.47	0.42	0.24	0.48	0.49	0.49	0.52	**0.54**
SSIM	0.26	0.22	0.22	0.30	0.19	0.24	0.23	0.26	0.26	**0.31**
MI	2.39	2.60	2.77	**2.98**	1.97	2.76	2.68	2.62	2.68	2.44
VIF	0.35	0.33	0.33	0.39	0.21	0.34	0.34	0.35	0.38	**0.40**
SCD	1.34	1.09	1.13	**1.45**	1.02	1.14	1.12	1.25	1.19	1.39

Table 3: Quantitative results of generalization experiment.

Generalization Experiment

Besides, we conduct generalization experiments on the M3SVD dataset (Tang et al. 2025), which contains real-captured degraded data using the MAG64AI camera. As shown in Fig. 10 and Table 3, RFC outperforms others in both visual quality and objective scores, demonstrating its strong generalization ability across diverse fusion scenarios.

Ablation Studies

Language-Feature Alignment Loss. $\mathcal{L}_{ali}$ is directly removed for verifying its role. As shown in Fig. 11, removing $\mathcal{L}_{ali}$ hinders full alignment with language instructions, leaving artifacts due to incomplete degradation removal. This can be attributed to the facilitating effect of $\mathcal{L}_{ali}$ on the reconstruction loss $\mathcal{L}_{rec}$. Specifically, $\mathcal{L}_{ali}$ aligns internal feature changes with the instruction, while $\mathcal{L}_{rec}$ enforces pixel-level consistency with the instruction-specified appearance. With the shared goal of ensuring the output faithfully follows the instruction, $\mathcal{L}_{ali}$ can help $\mathcal{L}_{rec}$ achieve better optimization, as shown in Fig. 11. The quantitative results in Table 4 also demonstrate the importance of $\mathcal{L}_{ali}$.

Functional and Spatial Conditions. We remove functional condition α and spatial condition β from the FiLM-inspired

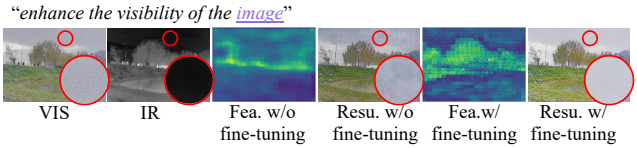

Figure 13: Ablation on fine-tuning CLIPSeg.

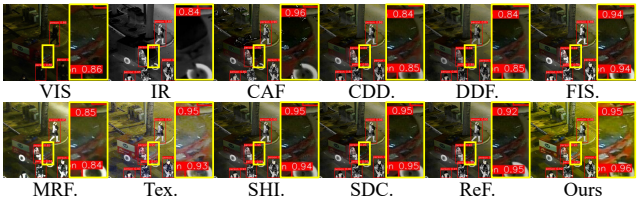

Figure 14: Qualitative object detection verification.

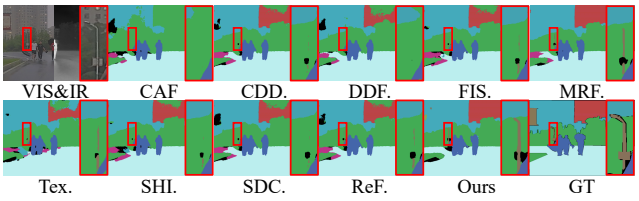

Figure 15: Qualitative semantic segmentation verification.

MCC module, respectively. As shown in Fig. 12. When removing α, RFC loses the ability to perform the specified degradation removal function. When removing β, RFC cannot specify regions for target processing. In contrast, RFC responds correctly only when both α and β are present. At this time, RFC is capable of accurately localizing the specified spatial regions while achieving the desired function of degradation removal.

Fine-Tuning CLIPSeg. Fine-tuning is applied to the decoder's last convolutional layers in CLIPSeg to adapt to multi-modal data, thereby improving both local localization and global perception. As shown in Fig. 13, fine-tuning enables image-wide responses and effective removal of global degradations, which are not present before fine-tuning. Additionally, the quantitative results in Table 4 show that removing fine-tuning leads to a decline in fusion performance.

The Number of Modules. In this work, we use three MCC modules to generate variables that capture both functional and spatial conditions, and use four HAF modules to embed the composite control priori into the fusion process. To verify the rationality of this configuration, we conduct an ablation study on the number of modules. As shown in Table 4, we evaluate the performance under different settings, using 2, 3, and 4 MCC modules and 3, 4, and 5 HAF modules, respectively. Our method achieves the best fusion performance with the configuration of 3 MCC & 4 HAF modules.

Semantic Verification on High-Level Tasks

Object Detection. We implement object detection on the LLVIP dataset with YOLO-v5, in which the detector is retrained on the results of these fusion methods and the source images. As shown in Fig. 14 and Table 5, our RFC surpasses all competitors in detection accuracy, demonstrating its ability to improve high-level semantic tasks. Notably, detection performance based on the fused images outperforms those based on the source images, highlighting the value of the image fusion technology.

Semantic Segmentation. We retrain SegFormer (Xie et al. 2021) on the FMB dataset and apply it to the fused results of each method. As shown in Fig. 15 and Table 6, RFC achieves superior segmentation performance across multiple categories, with results second only to MRFS. This is because MRFS is a specific method designed for the collaboration of image fusion and semantic segmentation.

Ablation	SSIM	MI	VIF	SCD	Qabf
w/o $\mathcal{L}_{ali}$	0.324	2.815	0.484	1.242	0.465
w/o Fine-tuning	0.321	2.817	0.479	1.226	0.460
2 MCC & 4 HAF	0.320	2.749	0.476	1.220	0.431
4 MCC & 4 HAF	0.323	2.687	0.474	1.172	0.439
3 HAF & 3 MCC	0.319	2.725	0.481	1.197	0.440
5 HAF & 3 MCC	0.324	2.732	**0.487**	1.193	0.435
Full Model (3 MCC & 4 HAF)	**0.325**	**2.867**	0.485	**1.242**	**0.469**

Table 4: Quantitative results of ablation studies.

LLVIP	Precision	Recall	mAP@0.6	mAP@0.85	mAP@(0.5:0.95)
VIS	79.0	62.5	65.4	50.6	53.5
IR	90.2	78.1	71.2	51.0	54.7
DDFM	95.6	85.8	85.4	61.7	71.1
SHIP	93.0	87.1	72.2	55.8	61.5
CDDFuse	93.9	89.0	80.6	52.8	65.5
MRFS	94.1	82.8	83.3	60.7	74.4
CAF	94.7	86.2	79.2	57.7	65.9
ReFusion	96.0	91.0	77.8	55.0	60.7
FISCNet	93.6	89.0	72.2	52.4	56.8
SDCFusion	93.2	88.6	72.2	53.5	56.4
Text-IF	96.0	90.9	79.5	58.3	64.8
Ours	96.7	90.2	87.5	61.5	**75.3**

Table 5: Quantitative object detection verification.

FMB	DDF.	SHI.	CDD.	MRF.	CAF	ReF.	FIS.	SDC.	Tex.	Ours
Vegetation	42.12	48.67	49.28	73.16	40.83	50.32	50.89	49.76	43.64	75.18
Building	56.81	62.01	60.18	78.7	55.42	62.4	60.73	59.27	53.24	81.38
Person	50.25	48.67	42.69	52.45	41.03	42.77	31.7	33.02	34.37	49.86
Car	70.63	75.07	72.61	77.66	70.21	74.66	73.91	72.59	70.71	77.59
Sky	54.39	67.34	69.28	90.5	50.3	69.94	72.3	71.13	58.19	89.6
mIoU	52.15	56.37	56.61	**67.54**	50.64	57.79	57.24	57.65	51.79	67.01

Table 6: Quantitative semantic segmentation verification.

Conclusion

This study proposes a robust fusion controller, termed RFC. It can achieve degradation-aware image fusion with fine-grained language instructions, improving the fusion model's robustness in harsh environments with spatial-varying composite degradations. RFC parses language instructions into functional and spatial conditions, and couples them to obtain the composite control priori. With the continuous modulation of this priori on the fusion process, combined with the guidance of the language-feature alignment loss, RFC can eliminate composite degradations according to language instructions. Extensive experiments demonstrate RFC's superiority in both perceptual performance and semantic quality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62506268, 62276192), the Fundamental Research Funds for the Central Universities (2042024kf0038), the Natural Science Foundation of Jiangsu Province (BK20250454), and the Postdoctoral Fellowship Program of CPSF (GZB20250066).

References

Bai, H.; Zhao, Z.; Zhang, J.; Wu, Y.; Deng, L.; Cui, Y.; Jiang, B.; and Xu, S. 2024. ReFusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *International Journal of Computer Vision*, 1: 1–21.

Chen, J.; Yang, L.; Liu, W.; Tian, X.; and Ma, J. 2024. Lenfusion: A joint low-light enhancement and fusion network for nighttime infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 73: 5018715.

Conde, M. V.; Geigle, G.; and Timofte, R. 2024. Instructir: High-quality image restoration following human instructions. In *Proceedings of the European Conference on Computer Vision*, 1–21.

Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; and Harada, T. 2017. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5108–5115.

Huang, Q.; Wu, G.; Jiang, Z.; Fan, W.; Xu, B.; and Liu, J. 2024. Leveraging a self-adaptive mean teacher model for semi-supervised multi-exposure image fusion. *Information Fusion*, 112: 102534.

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

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

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.; Lin, R.; Wu, G.; Liu, R.; Luo, Z.; and Fan, X. 2024a. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132(5): 1748–1775.

Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8115–8124.

Liu, J.; Wu, G.; Liu, Z.; Ma, L.; Liu, R.; and Fan, X. 2024b. Where elegance meets precision: Towards a compact, automatic, and flexible framework for multi-modality image fusion and applications. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1110–1118.

Liu, J.; Wu, G.; Liu, Z.; Wang, D.; Jiang, Z.; Ma, L.; Zhong, W.; and Fan, X. 2024c. Infrared and Visible Image Ffusion: From Data Compatibility to Task Adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 2349–2369.

Liu, R.; Liu, Z.; Liu, J.; Fan, X.; and Luo, Z. 2024d. A task-guided, implicitly-searched and metainitialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6594–6609.

Liu, X.; Huo, H.; Li, J.; Pang, S.; and Zheng, B. 2024e. A semantic-driven coupled network for infrared and visible image fusion. *Information Fusion*, 108: 102352.

Lüddecke, T.; and Ecker, A. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7086–7096.

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

Muller, A. C.; and Narayanan, S. 2009. Cognitively-engineered multisensor image fusion for military applications. *Information Fusion*, 10: 137–149.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3942–3950.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Singh, S.; Singh, H.; Bueno, G.; Deniz, O.; Singh, S.; Monga, H.; Hrisheekesha, P.; and Pedraza, A. 2023. A review of image fusion: Methods, applications and performance metrics. *Digital Signal Processing*, 137: 104020.

Tang, L.; Wang, Y.; Gong, M.; Li, Z.; Deng, Y.; Yi, X.; Li, C.; Xu, H.; Zhang, H.; and Ma, J. 2025. VideoFusion: A spatio-temporal collaborative network for mutli-modal video fusion and restoration. *arXiv preprint arXiv:2503.23359*.

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.

Tang, L.; Yuan, J.; and Ma, J. 2022. 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.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xu, H.; Ma, J.; Le, Z.; Jiang, J.; and Guo, X. 2020. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12484–12491.

Yadav, R.; Samir, A.; Rashed, H.; Yogamani, S.; and Dahyot, R. 2020. Cnn based color and thermal image fusion for object detection in automated driving. *Irish Machine Vision and Image Processing*, 2: 1–8.

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.

Zhang, H.; Tang, L.; Xiang, X.; Zuo, X.; and Ma, J. 2024a. Dispel darkness for better fusion: A controllable visual enhancer based on cross-modal conditional adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26487–26496.

Zhang, H.; Zuo, X.; Jiang, J.; Guo, C.; and Ma, J. 2024b. MRFS: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26974–26983.

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

Zhang, X.; Wang, X.; Yan, C.; and Sun, Q. 2024c. EV-fusion: A novel infrared and low-light color visible image fusion network integrating unsupervised visible image enhancement. *IEEE Sensors Journal*, 24: 4920–4934.

Zhao, Z.; Bai, H.; Zhang, J.; Zhang, Y.; Xu, S.; Lin, Z.; Timofte, R.; and Van Gool, L. 2023a. 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.; Zhu, Y.; Zhang, J.; Xu, S.; Zhang, Y.; Zhang, K.; Meng, D.; Timofte, R.; and Van Gool, L. 2023b. DDFM: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8082–8093.

Zheng, N.; Zhou, M.; Huang, J.; Hou, J.; Li, H.; Xu, Y.; and Zhao, F. 2024a. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26384–26395.

Zheng, N.; Zhou, M.; Huang, J.; and Zhao, F. 2024b. Frequency integration and spatial compensation network for infrared and visible image fusion. *Information Fusion*, 109: 102359.

Zou, D.; and Yang, B. 2023. Infrared and low-light visible image fusion based on hybrid multiscale decomposition and adaptive light adjustment. *Optics and Lasers in Engineering*, 160: 107268.