One-stop Multi-modality Image Dehazing-Registration-Fusion Framework via Progressive Task Cooperation*

Jing Li[1] , **Peiqi Cao**[2] and **Bin Yang**[3†]
[1]East China Normal University
[2]Beijing University of Posts and Telecommunicationsn
[3]Hunan University
jingli@geoai.ecnu.edu.cn, 2022312154@email.cufe.edu.cn, binyang@hnu.edu.cn

Abstract

Infrared and visible image fusion (IVF) encounters two challenges: 1) image unregistered, where parallax leads to blurred fusion results, and 2) adverse imaging conditions, which introduces outlier data (such as haze) that significantly degrade image registered and fusion performance. Existing IVF methods fail to account for the joint impact of unregistered and haze factor, often neglecting these challenges or addressing them in isolation. However, unregistered and haze are intrinsically coupled in fusion task. For example, haze can obscure structural details, leading to inaccurate registration and ultimately degrading fusion performance. To address the intrinsic coupling between unregistered and haze, we propose a progressive task-cooperative processing pipeline—dehazing, registration, and fusion—to achieve robust fusion of unregistered images under hazy conditions. To mitigate the challenges posed by architectural complexity and excessive parameter overhead in task progressive learning, we propose a lightweight model approximation paradigm through hierarchical knowledge distillation. The framework employs a stage-wise distillation optimization strategy that synergistically integrates: 1) primary task-specific distillation for modality-aware feature extraction, and 2) progressive task-cooperative distillation for fusion-oriented representation learning, which can improve fusion robustness for unregistered multi-modal inputs in hazy conditions. Extensive experiments demonstrate that our method achieves significantly superior performance compared to the State-Of-The-Art (SOTA) methods.

1 Introduction

IVF is a scene enhancement technique that leverages the complementary information provided by multiple sensors [Ma *et al.*, 2019a; Ma *et al.*, 2023]. Infrared sensors generate images by detecting the thermal radiation emitted by objects. Due to their longer wavelengths, they are less susceptible to degradation under adverse conditions such as haze [Bao *et al.*, 2023]. However, they exhibit limited sensitivity to fine-grained texture details. In contrast, visible sensors capture images based on the reflected light from objects, preserving rich texture information. Therefore, IVF integrates the complementary advantages of both sensors to achieve a more comprehensive representation of the scene, which have been widely applied in assisted driving and intelligent surveillance.

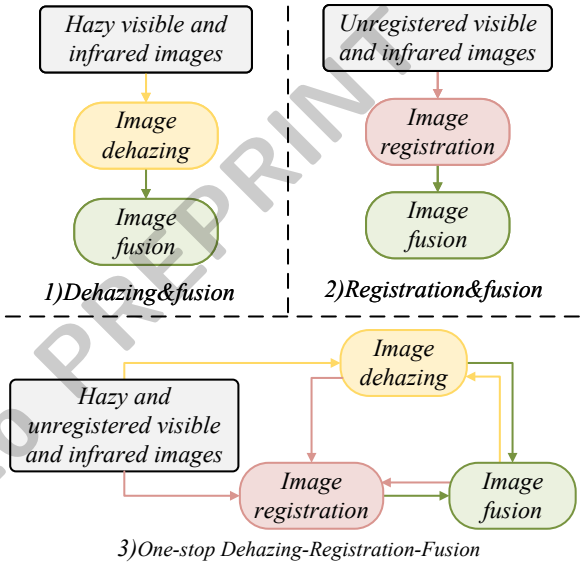

Figure 1: Differences of the existing and the proposed methods.

However, under complex condition like haze, they still face challenges in fusing hazy and unregistered images simultaneously. There exists a three-stage cascaded and tightly coupled relationship among dehazing, image registration, and fusion: **1) Dehazing affects the fusion performance.** Scattering caused by haze degrades texture in visible image, which in turn compromises the quality of fused image [He *et al.*, 2010; Cai *et al.*, 2016; Song *et al.*, 2023]. **2) Dehazing degrades registration accuracy.** Haze diminishes the visibility of fine image details, challenging discriminative feature extraction from degraded images and thus complicating feature-based registration. **3) Registration errors lead to fusion artifacts.** Imprecise registration introduces gradient redundancy

*Full version: https://github.com/null0315/full-version
†Corresponding author.

during image fusion, resulting in suboptimal visual quality [Wang *et al.*, 2022a; Xu *et al.*, 2023; Hong *et al.*, 2024; Li *et al.*, 2025a].

Although significant advances have been achieved in single tasks including dehazing, registration, and fusion, cascaded approaches are prone to semantic gaps and the accumulation of errors, such as Fig.1 (1-2). Existing researches have explored task collaboration in unregistered image fusion but has yet to address the joint optimization of multiple tasks (dehazing, registration, and fusion). To this end, we question: *Can a one-stop framework be designed to jointly accomplish dehazing, registration, and fusion by leveraging their task coupling?*

To answer this question, the following three key issues should be resolved. **1) Complexity of the multi-tasks.** The intertwining of dehazing, registration, and fusion task leads to feature discontinuities and the accumulation of error. **2) Large parameter size and optimization challenges.** The high parameter complexity of multi-task networks results in challenging convergence during training. **3) Robustness barrier.** The model performs well on simulated data but exhibits insufficient generalization in real-world scenarios.

Therefore, we propose a staged progressive training strategy that achieves joint multi-task optimization through hierarchical task interactions. In addition, we also introduce a multi-stage knowledge distillation [Gou *et al.*, 2021; Liu *et al.*, 2020] approach that employs knowledge transfer to build a lightweight end-to-end fusion framework. The main contributions of this work are summarized as follows:

- To address the intertwined challenges of dehazing, registration, and fusion in hazy unregistered multi-modal images. we propose the first unified Dehaze-Register-Fusion framework (DRF^2).
- To tackle inter-task coupling, we propose a phased progressive training strategy that mitigates feature-semantic gaps and error propagation across tasks.
- To address the large parameters and convergence issues in multi-task networks, we propose a novel pre-distillation and progressive distillation strategy for lightweight network construction.
- To validate the generalization capability of our method, we conduct cross-domain evaluations on both synthetic and real-world datasets, achieving better performance over SOTA methods.

2 Related Work

2.1 General Infrared and Visible Image Fusion

General IVF methods primarily address well-registered and non-degraded source images. With the advancement of deep learning, IVF has evolved from traditional methods based on handcrafted feature extraction and rule-based fusion to end-to-end approaches. According to the network, deep learning-based fusion methods include CNN- and GAN-based [Ma *et al.*, 2019b; Li and Wu, 2018] approaches for local feature extraction, Transformer-based [Wang *et al.*, 2022b] methods for global representation learning, and diffusion-based methods [Zhao *et al.*, 2023b] for denoising-oriented fusion.

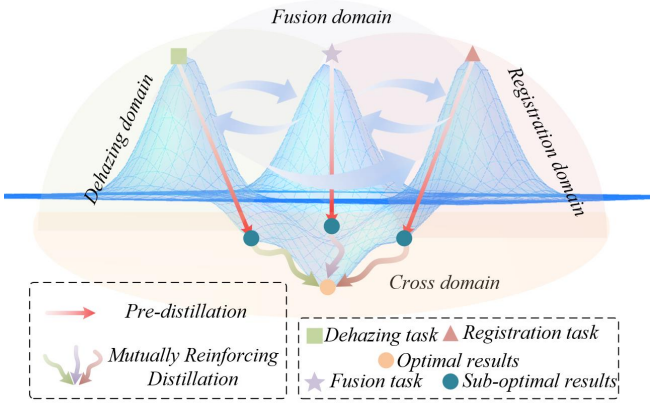

Figure 2: Stage-wise Distillation Optimization Strategy

To enhance the applicability of fusion results, mutually reinforcing methods have been developed, encompassing fusion guided by semantic segmentation or object detection [Liu *et al.*, 2023; Wu *et al.*, 2025], along with methods addressing unregistered image fusion [Wang *et al.*, 2022a; Hu *et al.*, 2025]. Recent text-guided IVF methods [Zhao *et al.*, 2024b] can enhance the flexibility of fusion but still fail to address the challenges arising from the coexistence of haze and unregistered images.

2.2 Unregistered Image Fusion

To address fusion blurring caused by registration errors and the limitations of the traditional register-then-fuse strategy, several unregistered image fusion methods are proposed. Xu et al. proposed MURF [Xu *et al.*, 2023] framework to solve unregistered IVF and medical image fusion by leveraging a mutually reinforcing mechanism between image registration and fusion. Li et al. proposed the MulFS-CAP [Li *et al.*, 2025a] framework to address the unregistered IVF. Hong et al. developed mutual-guided registration and fusion MERF [Hong *et al.*, 2024] model, which achieves multi-exposure image fusion through a mutual guidance mechanism. In addition, Wang et al. constructed a registration-fusion-segmentation framework to enable joint optimization of multiple tasks [Wang *et al.*, 2025].

2.3 Degraded Image Fusion

In complex weather conditions, atmospheric particle scattering (e.g., haze) leads to degradation in visible images, such as texture attenuation. To address this, various degradation-aware image fusion methods have been proposed. The Text-IF [Yi *et al.*, 2024] framework developed by Yi et al. introduces a text-guided mechanism to effectively handle compound degradation issues such as low light, overexposure, and low contrast. The OmniFuse [Zhang *et al.*, 2025] method proposed by Zhang et al. leverages large language models (LLMs) to achieve adaptive fusion across diverse degradation conditions. Meanwhile, MMAIF [Cao *et al.*, 2025] further integrates LLMs with diffusion models to tackle degradation-aware fusion tasks.

However, existing methods fall short in handling co-existing image unregistration and degradations like haze.

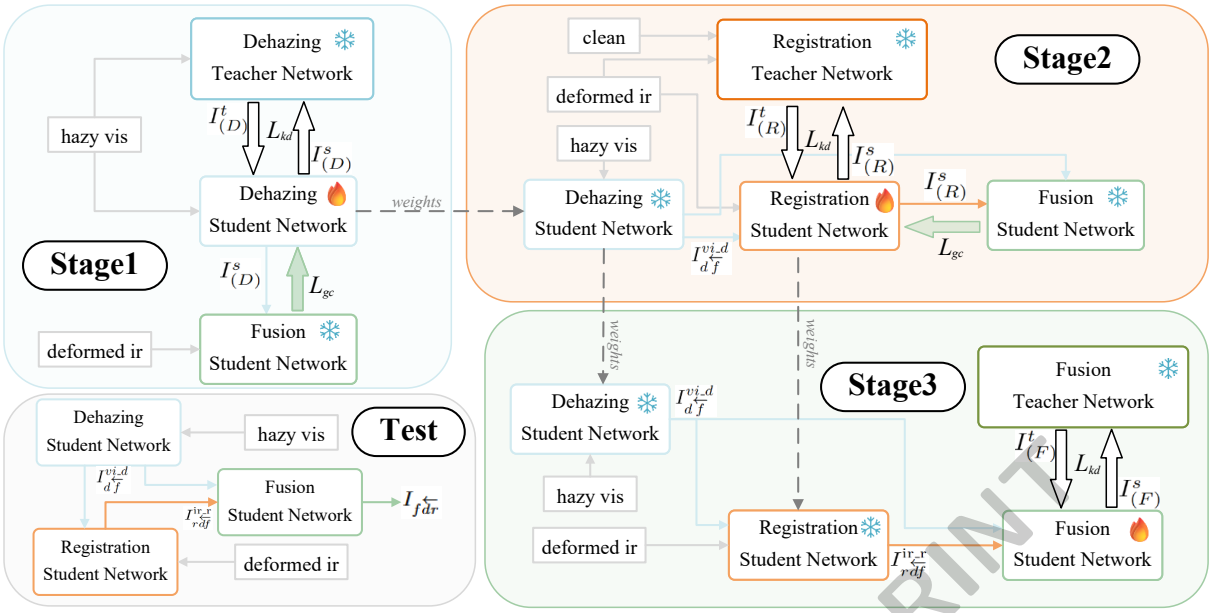

Figure 3: The architecture of the proposed method (Note: Prior to the three-stage progressive training, each student network is individually distilled from its teacher.)

3 Methodology

We propose a stage-wise distillation optimization strategy, as shown in Fig. 2 and Fig. 3. First, task-specific pre-distillation is employed to construct preliminary student networks for dehazing, registration, and fusion. Then, a progressive task-cooperative distillation mechanism is introduced—fusion facilitates dehazing; dehazing and fusion facilitate registration; dehazing and registration further enhance fusion—ultimately enabling end-to-end unregistered fusion of hazy images.

3.1 Problem Formulation

Conventional IVF methods assume that visible image is degradation-free and the source images are perfectly registered. The model can thus be formulated as:

$$I_f = G((I_{vi}, I_{ir}), \theta), \tag{1}$$

where I_{vi}, I_{ir}, I_f denote visible image, infrared image and fused image, $G(\cdot)$ denotes the mapping function of the fusion model, and θ represents the model parameters. Considering the hazy condition and the challenge of unregistered inputs, the mathematical formulation can be expressed as:

$$I_f = F((\hat{I}_{vi}, \tilde{I}_{ir}), \theta'), \tag{2}$$

where $\hat{I}_{vi}$ and $\tilde{I}_{ir}$ denote the hazy visible image and deformed infrared image, respectively, $F(\cdot)$ is the mapping function, and θ' denotes the model parameters.

3.2 Task-specific Pre-distillation

To address the challenges of joint optimization in unregistered hazy image fusion, we propose task-specific pre-distillation. It imparts prior knowledge to improve progressive task-cooperative distillation, mitigating suboptimal convergence caused by divergent task objectives. We design lightweight student networks by distilling knowledge from task-specific teachers. For dehazing, the ViT-based teacher [Song *et al.*, 2023] guides a 3-stage U-Net student with reduced channels, embedding dimensions, and attention ratios. For registration, the unsupervised cross-modality teacher [Wang *et al.*, 2022a] informs a student with a simplified feature extractor and a single lightweight module. For fusion, the high-performance equivariant teacher [Zhao *et al.*, 2024a] is distilled into a student with shallower pseudo-sensing and U-Fuser modules, omitting the costly equivariant augmentation.

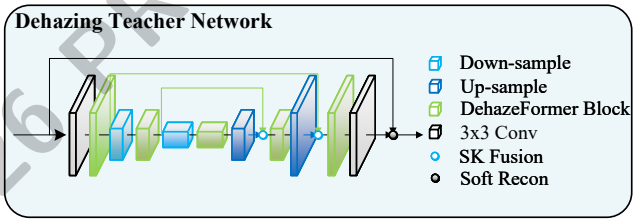

Figure 4: Dehazing Teacher Network

For task-specific pre-distillation, we apply a pixel-wise content loss L_{cont} between teacher and student outputs to preserve structural detail in image-level regression. L_{cont} is defined as follows:

$$L_{cont}(I^t_{(T)}, I^s_{(T)}) = \frac{1}{N}\sum_{i=1}^{N}\left\| I^t_{(T)i} - I^s_{(T)i} \right\|_1, \quad T \in \{D, R, F\}, \tag{3}$$

where $I^t_{(T)}$ and $I^s_{(T)}$ denote the outputs of the tasks, and N is

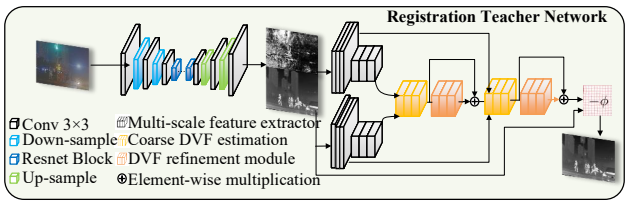

Figure 5: Registration Teacher Network

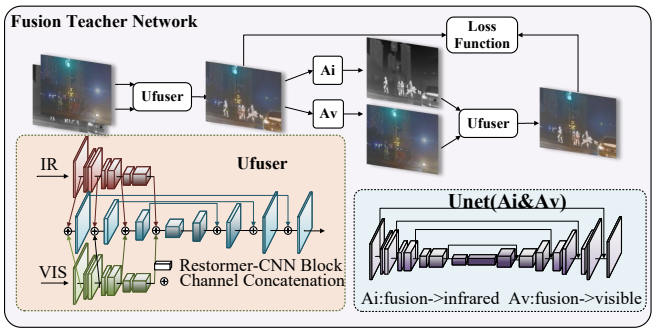

Figure 6: Fusion Teacher Network

the number of pixels in the image.

The perceptual loss L_{per} measures the semantic consistency between deep feature representations extracted by a pretrained network ϕ [Simonyan and Zisserman, 2014]:

$$L_{per}(I^t_{(T)}, I^s_{(T)}) = \frac{1}{C_j H_j W_j} \left\| \phi_j(I^t_{(T)}) - \phi_j(I^s_{(T)}) \right\|_2^2, \tag{4}$$

where C_j, H_j, and W_j denote the channel, height, and width of the j-th layer features, respectively.

The multi-scale loss L_{ms} enforces consistency across multiple spatial scales, improving robustness against local artifacts and hierarchical feature fidelity:

$$L_{\text{ms}}(I^t_{(T)}, I^s_{(T)}) = \sum_k \left\| F_k(I^t_{(T)}) - F_k(I^s_{(T)}) \right\|_1, \quad k \in \{1, 0.5, 0.25\}, \tag{5}$$

where F_k denotes downsampling to scale k.

In task-specific pre-distillation, the overall knowledge distillation loss L_{kd} combines the above losses:

$$L_{kd} = \alpha L_{cont} + \beta L_{per} + \gamma L_{ms}, \tag{6}$$

where α, β and γ are the weighting coefficients.

3.3 Progressive Task-Cooperative Distillation

We introduce progressive task-cooperative distillation, a stage-wise strategy that establishes a hierarchy of task dependencies to enable task cooperation, thereby addressing the coupled optimization of dehazing, registration, and fusion more effectively than cascaded approaches.

Fusion-Reinforced Dehazing

Haze degrades fusion performance, thus we employ the pre-distilled fusion network with fixed parameters, leveraging its learned fusion priors to enhance dehazing. The formulation is as follows:

$$I^{vi\_d}_{d\overleftarrow{f}} = F_{d\overleftarrow{f}}\left((\hat{I}_{vi}, \tilde{I}_{ir}), \theta_{d\overleftarrow{f}}\right), \tag{7}$$

where $I^{vi\_d}_{d\overleftarrow{f}}$ denotes the haze-free visible image, $F_{d\overleftarrow{f}}(\cdot)$ is the mapping function, and $\theta_{d\overleftarrow{f}}$ represents the parameters of the dehazing model. In addition to the distillation constraint from the dehazing teacher network, we further propose a gradient consistency-based constraint L_{gc}, formulated as:

$$L_{gc}(I) = \|f(I) - \max(f(I_{vis}), f(I_{ir}))\|_1, \tag{8}$$

where I denotes the input image, $f(\cdot)$ is the gradient magnitude operator (Sobel filter), and $\max(\cdot)$ computes the element-wise absolute maximum between the gradient maps of the clean visible image I_{vis} and the infrared image I_{ir}.

The total loss L_d combines the dehazing distillation loss with the gradient consistency loss:

$$L_d = L^d_{kd} + \lambda L^d_{gc}, \tag{9}$$

where L^d_{kd} and L^d_{gc} denotes the dehazing distillation loss of L_{kd} and L_{gc}, and λ is a weighting coefficient.

Jointly Guided Registration

To overcome the dual challenges of haze-impaired feature matching and misalignment-induced fusion artifacts, we boost registration accuracy with a pre-trained dehazing network and a pre-distilled fusion network, formulated as:

$$I^{\text{ir\_r}}_{r\overleftarrow{df}} = F_{r\overleftarrow{df}}\left((\tilde{I}_{\text{ir}}, I^{\text{vi\_d}}_{d\overleftarrow{f}}), \theta_{r\overleftarrow{df}}\right), \tag{10}$$

where $I^{\text{ir\_r}}_{r\overleftarrow{df}}$ denotes the registered infrared image, $F_{r\overleftarrow{df}}(\cdot)$ is the mapping function with parameters $\theta_{r\overleftarrow{df}}$. The total loss L_r is formulated as:

$$L_r = L^r_{\text{kd}} + \mu L^r_{\text{gc}}, \tag{11}$$

Unified Fusion Refinement

To effectively preserve complementary information while suppressing haze and parallax, the pre-distilled fusion network is further optimized through knowledge distillation, guided by pre-trained dehazing/registration models and utilizing haze-free visible and registered infrared images.

$$I_{f\overleftarrow{dr}} = F_{f\overleftarrow{dr}}\left((I^{\text{vi\_d}}_{d\overleftarrow{f}}, I^{\text{ir\_r}}_{r\overleftarrow{df}}), \theta_{f\overleftarrow{dr}}\right), \tag{12}$$

where $I_{f\overleftarrow{dr}}$ denotes the final fused result, $F_{f\overleftarrow{dr}}(\cdot)$ is the fusion mapping function with parameters $\theta_{f\overleftarrow{dr}}$.

In test stage, a hazy visible image and an unregistered infrared image are fed into our model to produce a haze-free fused image without distortions or parallax artifacts.

4 Experiment

4.1 Experimental Configurations

Datasets: The proposed model is trained on the simulated dataset and evaluated for generalization on both simulated and real-world datasets. Based on the FMB dataset [Cui *et*

	Experiments	Methods	Input data (other methods)	Input data (our method)
Experiments on simulated dataset	Dehazing and Fusion Comparison Experiments Registration and Fusion Comparison Experiments Dehazing, Registration and Fusion Comparison Experiments	Cascaded SOTA dehazing + fusion methods Registration-based fusion method Cascaded SOTA dehazing + Registration-based fusion method	Registered IR + Hazy VIS Unregistered IR + Hazy VIS/Clean VIS Unregistered IR + Hazy VIS	Unregistered IR + Hazy VIS
Experiments on real-world dataset	Dehazing and Fusion Comparison Experiments Registration and Fusion Comparison Experiments	Cascaded SOTA dehazing + fusion methods Cascaded SOTA dehazing + Registration-based fusion method	Registered IR + Hazy VIS Unregistered IR + Hazy VIS	Unregistered IR + Hazy VIS

Table 1: Data utilization of the comparative experiment

Method (dehzing&fusion)		PSNR	AG	EI	FD
-		12.607	1.394	14.624	1.615
fsnet (TPAMI '23)	MFEIF	15.390	2.340	24.801	2.689
dehazeformer (TIP '23)	(TCSVT '21)	15.498	2.366	25.170	2.700
DFPIR (CVPR '25)		15.079	1.915	20.271	2.211
-		9.302	1.752	18.270	2.084
fsnet (TPAMI '23)	SwinFusion	14.629	3.671	38.820	4.284
dehazeformer (TIP '23)	(JAS '22)	14.471	3.728	39.109	4.408
DFPIR (CVPR '25)		13.442	2.878	30.168	3.433
-		15.418	2.282	23.691	2.713
fsnet (TPAMI '23)	SegMiF	12.486	3.537	37.308	4.137
dehazeformer (TIP '23)	(ICCV '23)	12.515	3.642	38.297	4.277
DFPIR (CVPR '25)		13.978	3.598	37.761	4.271
-		12.503	2.848	29.466	3.427
fsnet (TPAMI '23)	cddfuse	14.705	3.706	39.047	4.353
dehazeformer (TIP '23)	(CVPR '23)	14.624	3.659	38.270	4.354
DFPIR (CVPR '25)		14.233	3.215	33.487	3.866
-		12.531	2.862	29.637	3.515
fsnet (TPAMI '23)	CFNet	9.786	1.889	20.043	2.315
dehazeformer (TIP '23)	(PR '24)	14.646	3.520	37.712	4.098
DFPIR (CVPR '25)		13.697	2.776	29.765	3.272
-		14.576	1.833	19.586	2.056
fsnet (TPAMI '23)	MaeFuse	15.332	3.638	38.992	4.068
dehazeformer (TIP '23)	(TIP '25)	<u>16.460</u>	3.468	37.138	3.889
DFPIR (CVPR '25)		15.159	2.639	28.335	2.948
-		9.530	1.572	16.124	1.895
fsnet (TPAMI '23)	SAGE	14.753	<u>4.257</u>	<u>44.577</u>	<u>5.008</u>
dehazeformer (TIP '23)	(CVPR '25)	15.505	3.882	40.428	4.638
DFPIR (CVPR '25)		13.309	2.854	29.893	3.370
Ours		**16.533**	**4.334**	**45.483**	**5.211**

Table 2: Quantitative results of cascaded SOTA dehazing + fusion methods on simulated dataset. Best and second-best values are **highlighted** and underlined.

al., 2023], we apply the depth-anything model [Yang *et al.*, 2024] and the atmospheric scattering model [Zhang *et al.*, 2017] to simulate haze on visible images and introduce both rigid and non-rigid transformations to infrared images. This experiment utilizes a self-collected dataset of 135 real-world haze image pairs. Each pair is provided in both registered and unregistered versions.

4.2 Experiments on Simulated Dataset

We compare our method—both qualitatively and quantitatively—with three baselines: 1) cascaded SOTA dehazing + fusion methods, 2) registration-based fusion methods, and 3) IVF methods. In addition, **Table 1 shows the input images for various comparative strategies.** For quantitative analysis, we adopt four quantitative metrics: Peak Signal-to-Noise Ratio (PSNR), Average Gradient (AG), Edge Intensity (EI), and Figure Definition (FD).

Dehazing and Fusion Comparisons

To validate the effectiveness of our method, we compare it with: 1) direct fusion without dehazing, and 2) cascaded dehazing-then-fusion. Dehazing methods include FSNet [Cui *et al.*, 2023], DehazeFormer [Song *et al.*, 2023] and DFPIR [Tian *et al.*, 2025], and the fusion methods include MFEIF [Liu *et al.*, 2021], SwinFusion [Ma *et al.*, 2022], SegMiF [Liu *et al.*, 2023], cddfuse [Zhao *et al.*, 2023a], CFNet [Xing *et al.*, 2024], MaeFuse [Li *et al.*, 2025b] and SAGE [Wu *et al.*, 2025]. Table 2 and Fig.7 show that our method outperforms all baselines on all metrics, and has a better visual effect.

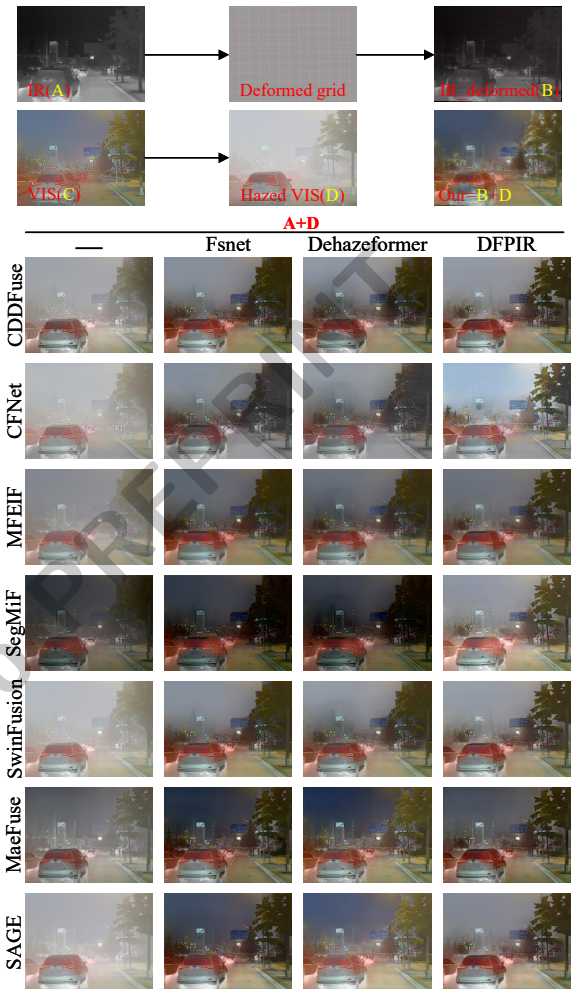

Figure 7: Qualitative results of cascaded SOTA dehazing + fusion methods on simulated dataset.

Method (registration&fusion)		PSNR	AG	EI	FD
UMF-CMGR (IJCAI '22)	clean	14.924	2.292	23.828	2.735
	hazy	14.884	1.562	16.549	1.816
MURF (TPAMI '23)	clean	15.783	<u>3.485</u>	<u>34.613</u>	<u>4.536</u>
	hazy	14.541	2.230	21.922	2.853
MulFS-CAP (TPAMI '25)	clean	<u>16.041</u>	3.122	33.127	3.617
	hazy	14.201	1.688	17.947	1.928
Ours		**16.533**	**4.334**	**45.483**	**5.211**

Table 3: Quantitative results of registration-based fusion method on simulated dataset.

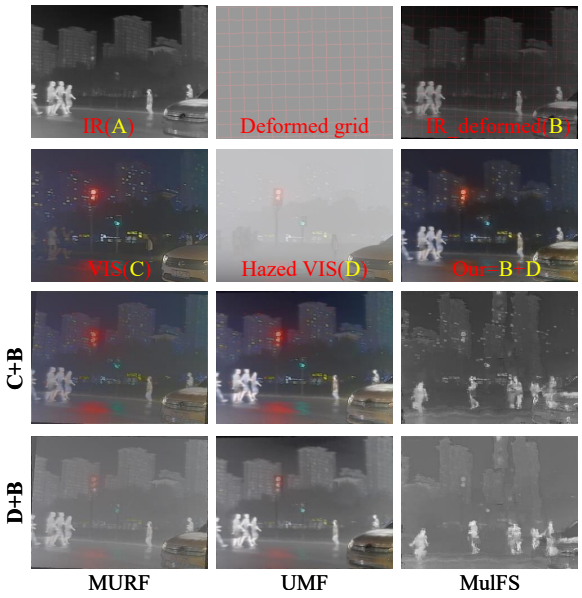

Figure 8: Qualitative results of registration-based fusion method on simulated dataset.

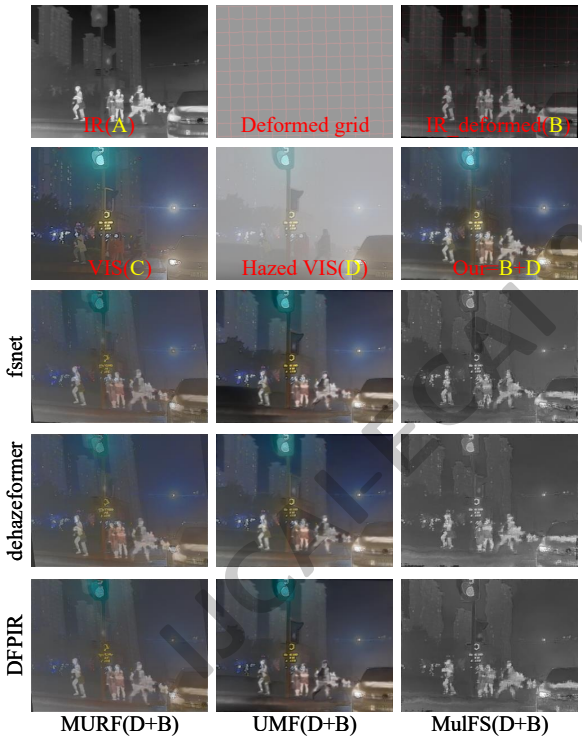

Figure 9: Qualitative results of cascaded SOTA dehazing + registration-based fusion method on simulated dataset.

Registration and Fusion Comparisons

This section focuses on unregistered image fusion and highlights the performance differences between the proposed method and registration-fusion coupled methods, which include UMF_CMGR [Wang *et al.*, 2022a], MURF [Xu *et al.*, 2023] and MulFS CAP [Li *et al.*, 2025a]. Table 3 and Fig.8 show that our method still achieves the best performance across all metrics while delivering superior visual quality, even under unequal comparison conditions.

Method (dehazing&registration&fusion)		PSNR	AG	EI	FD
fsnet (TPAMI '23)	UMF-CMGR (IJCAI '22)	15.902	2.905	30.730	3.410
dehazeformer (TIP '23)		15.703	2.717	28.551	3.257
DFPIR (CVPR '25)		15.204	2.415	25.700	2.810
fsnet (TPAMI '23)	MURF (TPAMI '23)	15.662	3.186	31.605	4.146
dehazeformer (TIP '23)		15.503	3.509	34.972	4.572
DFPIR (CVPR '25)		15.161	2.672	26.512	3.479
fsnet (TPAMI '23)	MulFS-CAP (TPAMI '25)	16.175	3.424	36.447	3.959
dehazeformer (TIP '23)		16.212	3.201	34.061	3.702
DFPIR (CVPR '25)		15.835	2.649	28.232	3.054
Ours		**16.533**	**4.334**	**45.483**	**5.211**

Table 4: Quantitative results of cascaded SOTA dehazing + registration-based fusion method on simulated dataset.

Method (dehazing&fusion)		EN	AG	RMSC	SD
-	MFEIF (TCSVT '21)	6.867	50.007	40.281	0.158
fsnet (TPAMI '23)		6.437	34.853	29.191	0.114
dehazeformer (TIP '23)		6.489	35.598	30.004	0.118
DFPIR (CVPR '25)		6.365	33.773	27.616	0.108
-	SwinFusion (JAS '22)	6.804	43.550	39.030	0.153
fsnet (TPAMI '23)		6.880	46.257	41.369	0.162
dehazeformer (TIP '23)		6.921	47.896	41.963	0.165
DFPIR (CVPR '25)		6.813	43.998	39.329	0.154
-	SegMiF (ICCV '23)	6.670	48.514	33.580	0.132
fsnet (TPAMI '23)		6.741	49.381	35.985	0.141
dehazeformer (TIP '23)		6.784	49.937	37.127	0.146
DFPIR (CVPR '25)		6.809	53.404	38.446	0.151
-	cddfuse (CVPR '23)	6.899	47.964	41.450	0.163
fsnet (TPAMI '23)		6.945	48.991	43.260	0.170
dehazeformer (TIP '23)		6.972	49.618	43.722	0.171
DFPIR (CVPR '25)		6.904	47.900	41.660	0.163
-	CFNet (PR '24)	6.851	51.273	38.951	0.153
fsnet (TPAMI '23)		6.919	53.749	41.139	0.161
dehazeformer (TIP '23)		6.896	52.916	40.878	0.160
DFPIR (CVPR '25)		6.967	57.155	40.619	0.159
-	MaeFuse (TIP '25)	6.680	49.340	28.804	0.113
fsnet (TPAMI '23)		6.779	51.178	30.624	0.120
dehazeformer (TIP '23)		6.927	55.317	35.055	0.137
DFPIR (CVPR '25)		6.684	49.475	29.086	0.114
-	MaeFuse (TIP '25)	6.710	46.171	34.887	0.137
fsnet (TPAMI '23)		6.948	51.555	40.747	0.160
dehazeformer (TIP '23)		**6.975**	57.281	40.787	0.160
DFPIR (CVPR '25)		6.776	46.946	35.788	0.140
Ours		6.964	**57.810**	**44.409**	**0.174**

Table 5: Quantitative results of cascaded SOTA dehazing + fusion methods on real dataset.

Dehazing, Registration and Fusion Comparisons

To evaluate the joint processing capability in complex scenes, we compare our method against multi-stage pipelines that sequentially apply SOTA dehazing (FSNet [Cui *et al.*, 2023], DehazeFormer [Song *et al.*, 2023] and DFPIR [Tian *et al.*, 2025]), registration-based fusion methods (UMF_CMGR [Wang *et al.*, 2022a], MURF [Xu *et al.*, 2023], and MulFS CAP [Li *et al.*, 2025a]). As shown in Table 4 and Fig. 9, our method achieves the best scores across all metrics.

4.3 Experiments on Real-world Dataset

Dehazing and Fusion Comparisons

To validate the generalization of our method, we conducted extended experiments on a real dataset comparing: 1) direct fusion without dehazing; 2) cascaded dehazing-fusion using the same modules as in above section. **Table 1 shows the input images for various comparative strategies.** Due to the absence of clean visible images as ground truth, we se-

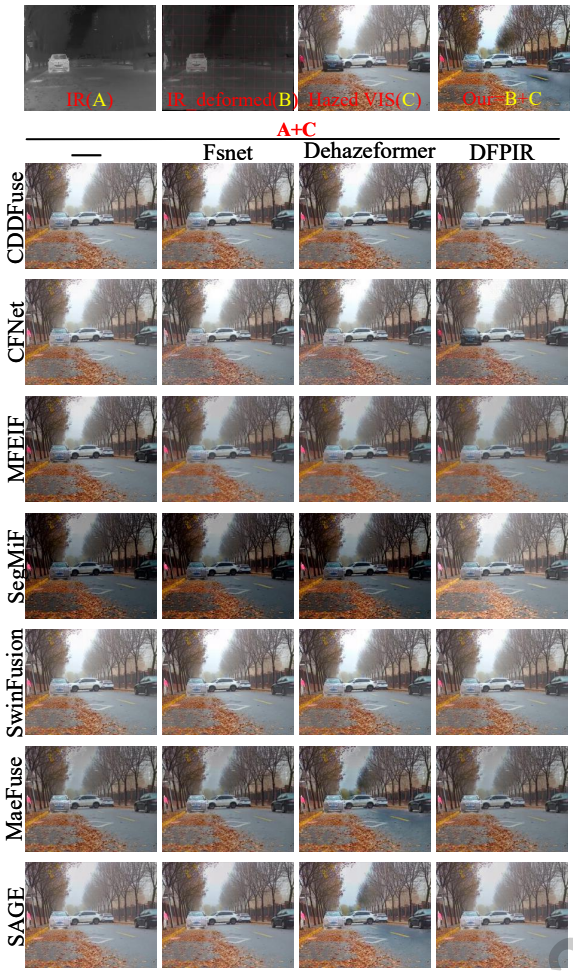

Figure 10: Qualitative results of cascaded SOTA dehazing + fusion methods on real dataset.

Method (dehazing&registration&fusion)		EN	AG	RMSC	SD
-	UMF-CMGR (IJCAI '22)	6.324	29.020	23.086	0.091
fsnet (TPAMI '23)		6.799	36.699	31.926	0.125
dehazeformer (TIP '23)		6.848	42.876	31.631	0.124
DFPIR (CVPR '25)		6.739	35.686	30.298	0.119
-	MURF (TPAMI '23)	6.252	33.834	23.877	0.094
fsnet (TPAMI '23)		6.366	45.069	23.325	0.091
dehazeformer (TIP '23)		6.337	47.843	22.322	0.088
DFPIR (CVPR '25)		6.340	44.332	22.779	0.089
-	MulFS-CAP (TPAMI '25)	**7.195**	47.317	43.845	0.172
fsnet (TPAMI '23)		6.709	51.012	30.807	0.121
dehazeformer (TIP '23)		6.711	54.662	30.208	0.118
DFPIR (CVPR '25)		6.631	52.331	28.942	0.113
Ours		6.964	**57.810**	**44.409**	**0.174**

Table 6: Quantitative results of cascaded SOTA dehazing + registration-based fusion method on real dataset.

lect four no-reference metrics for quantitative analysis, which includes entropy (EN), average gradient (AG), root mean square contrast (RMSC), and standard deviation (SD). Table 5 and Fig.10 show that *despite the unfair comparison conditions*, our method still achieves the best results on three metrics and delivers superior visual quality.

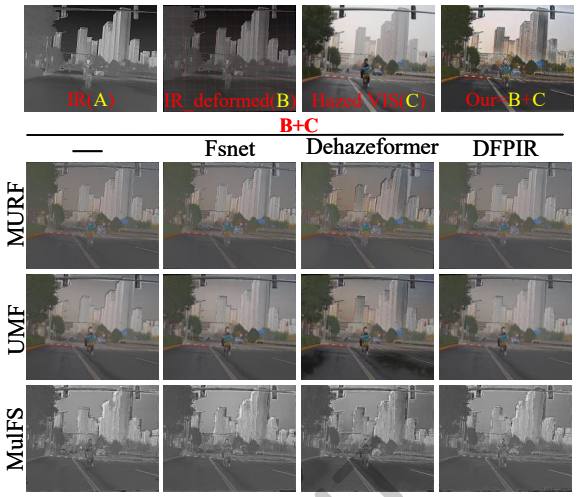

Figure 11: Qualitative results cascaded SOTA dehazing + registration-based fusion method on real dataset.

Dehazing, Registration and Fusion Comparisons

Consistent with the simulation study, we extend our evaluation to real-world environments, comparing our approach with the aforementioned multi-stage baselines. As shown in Table 6 and Fig.11, our method achieves the highest scores across three quantitative metrics. Our framework is free from the ghosting and blurring that plague cascaded approaches, achieving sharp fusion by simultaneously aligning and dehazing source images.

4.4 Parameter Analysis

Employing pre-distillation and progressive task-cooperative distillation, we construct lightweight networks to resolve the high-parameter and convergence challenges in multi-task settings. The student networks thus have 21.48 M fewer parameters than their teachers and are more parameter-efficient than other methods.

5 Conclusion

We propose DRF^2, a lightweight and robust IVF framework that explicitly addresses the intrinsic coupling between haze and unregistered through a cooperative dehazing-registration-fusion pipeline. By introducing a progressive task-cooperation mechanism with hierarchical knowledge distillation, our method enables bidirectional enhancement across tasks, significantly improving fusion quality under adverse conditions. Extensive experiments validate the advantages of DRF^2.

References

[Bao *et al.*, 2023] Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 619(7971):743–748, 2023.

[Cai *et al.*, 2016] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end

system for single image haze removal. *IEEE transactions on image processing*, 25(11):5187–5198, 2016.

[Cao *et al.*, 2025] Zihan Cao, Yu Zhong, Ziqi Wang, and Liang-Jian Deng. Mmaif: Multi-task and multi-degradation all-in-one for image fusion with language guidance. *arXiv preprint arXiv:2503.14944*, 2025.

[Cui *et al.*, 2023] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Image restoration via frequency selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1093–1108, 2023.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[He *et al.*, 2010] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[Hong *et al.*, 2024] Wenhui Hong, Hao Zhang, and Jiayi Ma. Merf: a practical hdr-like image generator via mutual-guided learning between multi-exposure registration and fusion. *IEEE Transactions on Image Processing*, 33:2361–2376, 2024.

[Hu *et al.*, 2025] Dengshu Hu, Ke Wang, Cuijin Zhang, Zheng Liu, Yukui Che, Shoubing Dong, and Chuirui Kong. Target-aware unregistered infrared and visible image fusion. *Frontiers in Physics*, 13:1599968, 2025.

[Li and Wu, 2018] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

[Li *et al.*, 2025a] Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Li *et al.*, 2025b] Jiayang Li, Junjun Jiang, Pengwei Liang, Jiayi Ma, and Liqiang Nie. Maefuse: Transferring omni features with pretrained masked autoencoders for infrared and visible image fusion via guided training. *IEEE Transactions on Image Processing*, 2025.

[Liu *et al.*, 2020] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020.

[Liu *et al.*, 2021] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):105–119, 2021.

[Liu *et al.*, 2023] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8115–8124, 2023.

[Ma *et al.*, 2019a] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019.

[Ma *et al.*, 2019b] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48:11–26, 2019.

[Ma *et al.*, 2022] 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 Journal of Automatica Sinica*, 9(7):1200–1217, 2022.

[Ma *et al.*, 2023] Weihong Ma, Kun Wang, Jiawei Li, Simon X Yang, Junfei Li, Lepeng Song, and Qifeng Li. Infrared and visible image fusion technology and application: A review. *Sensors*, 23(2):599, 2023.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Song *et al.*, 2023] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023.

[Tian *et al.*, 2025] Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28165–28175, 2025.

[Wang *et al.*, 2022a] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*, 2022.

[Wang *et al.*, 2022b] Zhishe Wang, Yanlin Chen, Wenyu Shao, Hui Li, and Lei Zhang. Swinfuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022.

[Wang *et al.*, 2025] Di Wang, Xianghao Jiao, Jinyuan Liu, and Xin Fan. Robust one-stop multi-modality image registration-fusion-segmentation framework against misalignments and adversarial attacks. *IEEE Transactions on Multimedia*, 2025.

[Wu *et al.*, 2025] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17882–17891, 2025.

[Xing *et al.*, 2024] Mengliang Xing, Gang Liu, Haojie Tang, Yao Qian, and Jun Zhang. Cfnet: An infrared and visible image compression fusion network. *Pattern Recognition*, 156:110774, 2024.

[Xu *et al.*, 2023] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and

fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):12148–12166, 2023.

[Yang *et al.*, 2024] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024.

[Yi *et al.*, 2024] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024.

[Zhang *et al.*, 2017] Yanfu Zhang, Li Ding, and Gaurav Sharma. Hazerd: an outdoor scene dataset and benchmark for single image dehazing. In *2017 IEEE international conference on image processing (ICIP)*, pages 3205–3209. IEEE, 2017.

[Zhang *et al.*, 2025] Hao Zhang, Lei Cao, Xuhui Zuo, Zhenfeng Shao, and Jiayi Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Zhao *et al.*, 2023a] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.

[Zhao *et al.*, 2023b] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023.

[Zhao *et al.*, 2024a] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024.

[Zhao *et al.*, 2024b] Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, et al. Image fusion via vision-language model. *arXiv preprint arXiv:2402.02235*, 2024.