SpeciFuse: Learning Degradation-Type Specificity for Robust Infrared and Visible Image Fusion Under Composite Degradations

Xuan Li[1] , **Zhaoming Feng**[1] , **Xiang Yuan**[1] , **Huabing Zhou**[1] , **Jiayi Ma**[2,*]
[1] Wuhan Institute of Technology
[2] Wuhan University
lixuan@wit.edu.cn, {22303010013, 22503010043}@stu.wit.edu.cn, {zhouhuabing, jyma2010}@gmail.com

Abstract

Existing degradation-resistant infrared-visible image fusion methods struggle to effectively handle composite degradations, where multiple degradation types exhibit intricate coupling and mutual interference. To address this challenge, we propose SpeciFuse, an infrared-visible image fusion network that learns degradation-specific representations. By explicitly modeling the specificity of individual degradations through a siamese architecture trained on single-degradation data, our method effectively addresses the conflicts arising from multiple degradations, thereby achieving robust fusion performance that generalizes to diverse composite degradation scenarios. In SpeciFuse, a degradation-type specificity decoupling module is designed to disentangle the feature representations by attenuating their redundant correlations in a latent space. It minimizes off-diagonal elements to enforce independence among the encodings of distinct degradation types, while preserving dominant diagonal elements to maintain shared content information. Furthermore, a degradation-aware gated fusion module leverages these decoupled features to dynamically modulate cross-modal fusion across both channel and spatial dimensions. This facilitates a degradation-conditional fusion process that adaptively suppresses degradation artifacts while preserving beneficial information across varying conditions. In extensive comparisons with SOTA methods, SpeciFuse demonstrates more robust performance under diverse composite degradations. The code is available at https://github.com/xbsj-cool/SpeciFuse.

1 Introduction

Infrared-visible image fusion (IVIF) is a prominent subfield in multimodal image fusion research [Zhang and Demiris, 2023]. While visible images provide texturally rich details consistent with human vision, they are vulnerable to poor lighting and camouflage. Infrared images highlight thermal targets and penetrate obscurants like smoke and darkness, but they typically suffer from low spatial resolution and are more susceptible to noise. Therefore, IVIF aims to generate a single image that combines these complementary information to achieve more robust visual perception [Ma *et al.*, 2019a]. Furthermore, the fused images significantly improve the performance of downstream computer vision tasks such as object detection, tracking, and segmentation [Li *et al.*, 2022; Chen *et al.*, 2025; Zhao *et al.*, 2024; Li *et al.*, 2025c].

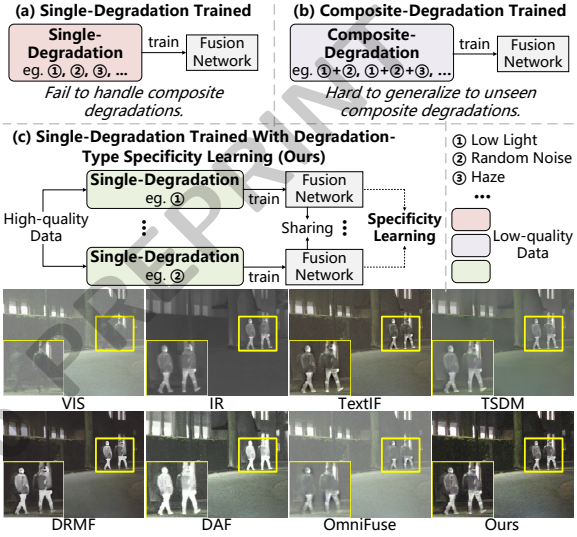

Figure 1: Existing degradation-resistance fusion approaches and their fusion results under composite degradation.

Although deep learning has revolutionized IVIF by offering superior performance, most existing methods assume ideal input conditions. Their performance thus significantly deteriorates when source images are degraded. This limitation has motivated growing interest in integrating image restoration with fusion processes [Guo *et al.*, 2024; Wang *et al.*, 2025] to enhance robustness in real-world scenarios, as shown in Figure 1. Single-degradation data refers to images that are corrupted by only one type of degradation (e.g., low light, blur or noise). Composite-degradation data consists of images that are affected by multiple degradation types. Methods trained on single-degradation data [Yi *et al.*, 2024; Tang *et al.*, 2025b] simply aggregate the capacity to address each isolated degradation, and thus fail to handle images with composite degradation. Methods trained on composite-

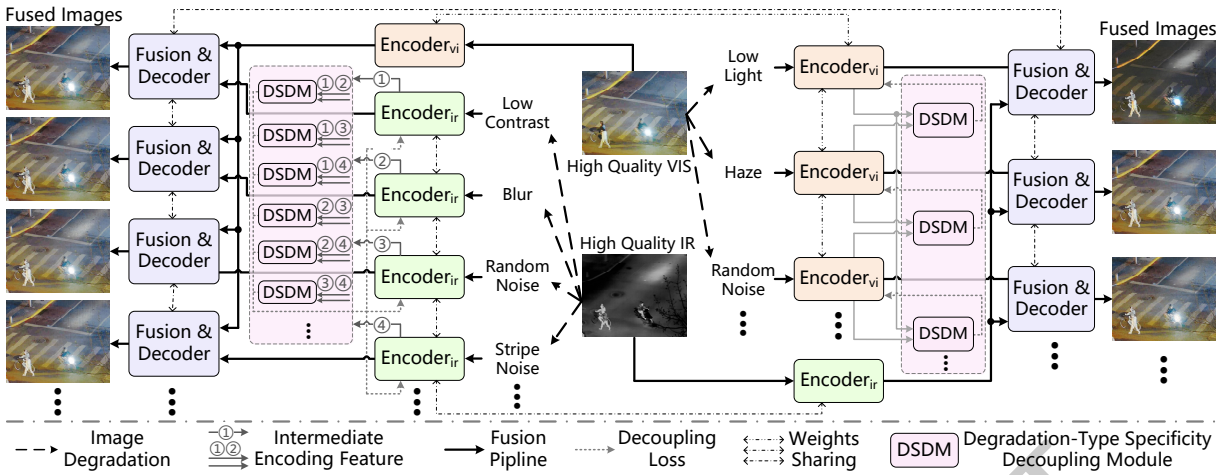

Figure 2: The siamese architecture of SpeciFuse.

degradation data [Xue *et al.*, 2024; Tang *et al.*, 2024; Zhang *et al.*, 2025] learn holistic mapping from fixed degradation combinations, resulting in limited generalization to unseen composite degradation scenarios. The neglect of the intricate coupling and mutual interference among distinct degradation types hinders aforementioned methods from discerning degradation types and correspondingly suppressing them in the fusion process, thereby inevitably leading to suboptimal fusion results under diverse composite degradations.

To overcome the above limitations, we propose SpeciFuse, an IVIF network designed to explicitly model and address conflicts arising from composite degradations, thereby achieving robust image fusion that generalizes effectively to various combinations of multiple degradations. As shown in Figure 2, SpeciFuse employs a siamese architecture that has multiple parallel fusion pipelines to learn the specificity of individual degradation type and to adaptively modulate fusion strategy. During training, degraded images are first generated from paired high-quality source images by using distinct single-degradation operations. Each fusion pipeline takes one degraded image and the high-quality image from the other modality as the input. All fusion pipelines share the same network architecture and weights, and their outputs are jointly optimized to train the SpeciFuse model.

To capture the specificity of individual degradations, we propose degradation-type specificity decoupling modules (DSDM), which establish interactions between the encoders of degraded modality in two pipelines, as shown in Figure 3. The DSDM projects the intermediate features of encoders into a latent space and computes their cross-correlation matrix. It constrains these matrices to an approximately diagonal form, where larger diagonal elements ensure the robust representation of shared content, and smaller off-diagonal elements compel each feature channel to specialize in encoding characteristics unique to individual degradation. Through this strategy, SpeciFuse learns to disentangle degradation-specific features, thereby distinguishing the existing types of degradation from intricate coupling while preserving content fidelity under diverse composite degradations.

Furthermore, we propose a degradation-aware gated fusion module (DGFM) to dynamically modulate the fusion process according to the decoupled degradation-type-specific features, as shown in Figure 3. Specifically, the DGFM first derives a degradation attention map from the encoded features. Capturing the characteristics of the prevailing degradation types, this map is then used to generate channel-wise and spatial-wise gating weights that directly modulate the fusion of cross-modal features. Through this degradation-conditional gating mechanism, SpeciFuse reliably generates high-quality fused images even under diverse and challenging degradation conditions. The contributions of this paper are summarized as follow:

- We propose SpeciFuse, an infrared-visible image fusion network that consists of multiple fusion pipelines based on a siamese architecture. It is designed to explicitly learn the specificity of individual degradations and adaptively modulate the fusion process, enabling robust generalization to composite degradation scenarios while being trained on single-degradation conditions.
- Degradation-type specificity decoupling modules are designed between two pipelines to drive the cross-correlation matrix of their encoded features toward a diagonal form. This strategy facilitates clear discernment of degradation-type-specific features from composite degradations, thus eliminating mutual interference while preserving content fidelity.
- A degradation-aware gated fusion module is proposed to dynamically modulate the fusion strategy in both channel and spatial dimensions based on the decoupled degradation-type-specific features. It establishes a dynamic fusion strategy that generates high-quality fusion results under diverse scenarios.

2 Related Works

2.1 Deep Learning-based Image Fusion Methods

With the advancement of deep learning, deep learning-based methods [Tang *et al.*, 2025a; Li *et al.*, 2025a; Liu *et al.*, 2024;

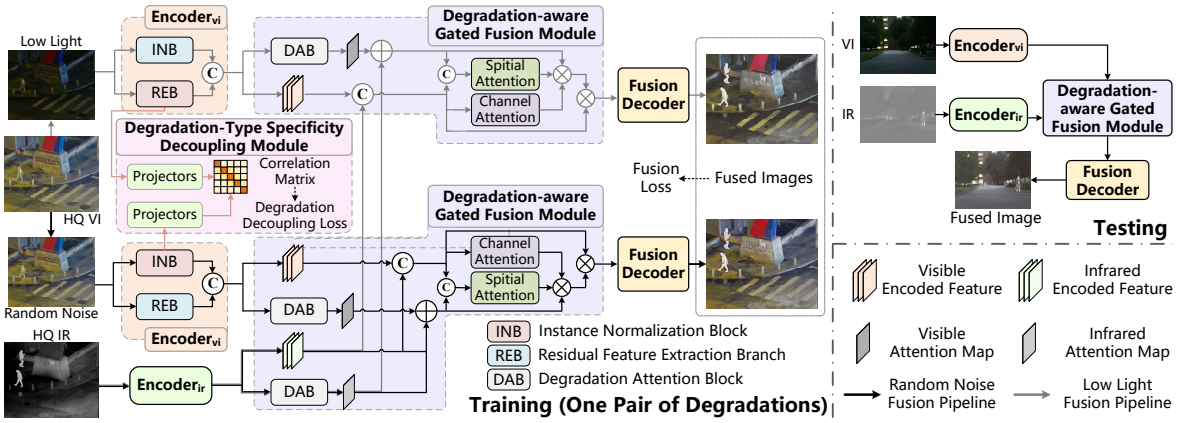

Figure 3: The structure detail of SpeciFuse. For clarity of demonstration, we only show the fusion pipeline of visible low-light and noise degradation in the training phase.

Tang *et al.*, 2025c; Li *et al.*, 2025b] have become the mainstream in IVIF. They can be broadly categorized into auto-encoder-based, GAN-based, and end-to-end approaches.

Auto-encoder-based methods typically employ an encoder-decoder framework. DenseFuse [Li and Wu, 2018] pioneers this direction by extracting multi-scale features, but its fusion strategy is oversimplified. Xu et al. [Xu *et al.*, 2021] later proposed a learnable adaptive fusion strategy to overcome handcrafted rules, though at increased computational cost. GAN-based methods formulate fusion as an adversarial game. FusionGAN [Ma *et al.*, 2019b] is a pioneering work in this field. Ma et al. [Ma *et al.*, 2020] introduced a dual-discriminator architecture to better balance multimodal information. However, these methods often suffer from training instability and mode collapse. End-to-end methods have evolved to implicitly handle the entire fusion pipeline. Zhang et al. [Zhang *et al.*, 2020] refined the loss function to guide information distribution. PIAFusion [Tang *et al.*, 2022a] introduced illumination-aware and differential-aware mechanisms. CDDFuse [Zhao *et al.*, 2023] proposed a correlation-driven dual-branch architecture. SwinFusion [Ma *et al.*, 2022] utilized a Swin Transformer to model long-range dependencies. YDTR [Tang *et al.*, 2022b] designed a Y-shape dynamic Transformer for multi-task generalization.

2.2 Degradation-Resistant Image Fusion

To address information loss in complex scenes, degradation-resistant IVIF methods have been developed. Initially, researches focused on low-light conditions. Lei et al. [Lei *et al.*, 2024] developed a method specifically for nighttime scenarios. DIVFusion [Tang *et al.*, 2023] jointly addresses low-light enhancement and fusion in an end-to-end framework. LENFusion [Chen *et al.*, 2024] integrates a feedback mechanism for joint low-light enhancement and fusion.

As real-world degradations extend beyond low-light, research shifted toward handling multiple degradations. TextIF [Yi *et al.*, 2024] employs a text-interactive approach to dynamically adjust fusion based on degradation descriptions. ControlFusion [Tang *et al.*, 2025b] integrates language-vision prompts for controllable and adaptive fusion.

More recent works aim for automated adaptation to composite degradations. TSDM [Xue *et al.*, 2024] employs a teacher-student framework using synthesized degraded images. DRMF [Tang *et al.*, 2024] leverages diffusion models to aggregate cross-modal priors during the restoration and fusion process. DAF [Wang *et al.*, 2025] learns degradation recovery implicitly by constructing dynamic negative and invariant positive samples. OmniFuse [Zhang *et al.*, 2025] couples restoration and fusion in a diffusion model and enables language control. However, the neglect of the intricate coupling and mutual interference among distinct degradation types hinders these methods from generate high-quality results under diverse composite degradations.

3 Method

The architecture of the proposed SpeciFuse is shown in Figure 2. It employs multiple fusion pipelines within a siamese architecture, each dedicated to processing images of one specific degradation. For example, in a pipeline processing a low-light degraded visible image, the corresponding infrared input remains in its original high-quality state. The total number of pipelines N is defined as the sum of the degradation types from both modalities (i.e., $N = N_{vi} + N_{ir}$). Each pipeline contains feature encoders for both modalities, a degradation-aware gated fusion module, and a decoder for reconstructing the fused image. The corresponding components share weights across all pipelines.

3.1 Degradation-Type Specificity Decoupling Module

Existing degradation-resistant image fusion methods suffer from limited generalization capabilities due to the intricate coupling and mutual interference among distinct degradations. To address this, we propose degradation-type specificity decoupling module (DSDM) to enforces the separation of feature representations from different types of degradation during encoding, thereby preventing feature interference that would otherwise compromise the fusion quality, as demonstrated in Figure 3.

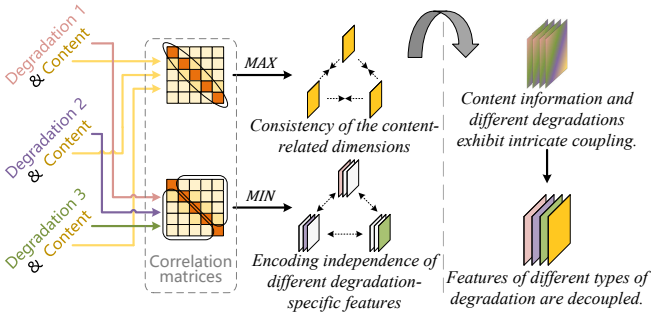

Figure 4: The explanation of degradation-type specificity decoupling. Degradation 1/2/3 are differently degraded images of the same source image, and the cross-correlation matrix is calculated between encoded features of two pipelines.

Inspired by the Barlow Twins [Zbontar *et al.*, 2021], we propose a novel degradation-type specificity decoupling module to address this fundamental limitation. It explicitly establishes interaction between differently degraded images with same content information, and guides the network to decouple degradation features by decorrelating their feature representations in the latent space (Figure 4). This strategy enforces each degradation type to be encoded in distinct feature dimensions. It enables clear discernment of degradations from mutual interference and preserves high-fidelity content within the encoded features, thus facilitating the capability of handling composite degradations.

Specifically, during the training phase, for a paired of differently degraded images derived in same modality, the output features of instance normalization [Huang and Belongie, 2017] blocks F_d are processed through two successive projection operations. The flattened features of the i-th degradation F^i_{flat} are obtained by two projectors and a convolution layer.

For a pair of flattened features of differently degraded images, they describe identical scene content but carry distinct types of degradation. Thus a degradation decoupling loss is computed based on the cross-correlation matrix of two flattened features. The diagonal elements are maximized to ensure consistent representation of objects and scene content across varying degradation conditions. Simultaneously, the off-diagonal elements are minimized to force each dimension to specifically encode features of particular degradation and thus obtain their distinctive features. The total degradation-type specificity decoupling loss is expressed as:

$$L_{SPE} = \sum\nolimits_{(i,j)} L_{spe} + \sum\nolimits_{(p,q)} L_{spe}, \quad (1)$$
$$i,j = 1,2,...,N_{vi}, i \neq j, \quad p,q = 1,2,...,N_{ir}, p \neq q,$$

where L_{spe} denote the decoupling loss in single modality. It can be described as:

$$L_{spe} = \sum\nolimits_m (1 - C_{m,m})^2 + \lambda \sum\nolimits_m \sum\nolimits_{n \neq m} C^2_{m,n}, \quad (2)$$

where i and j represent the i-th and j-th degradation, respectively. C is the cross-correlation matrix and $C_{m,m}$ denotes the matrix element at position (m, m). λ is a weighting hyperparameter that balances the contributions of the two loss components. We set $\lambda = 0.0051$, which is consistent with the Barlow Twins implementation.

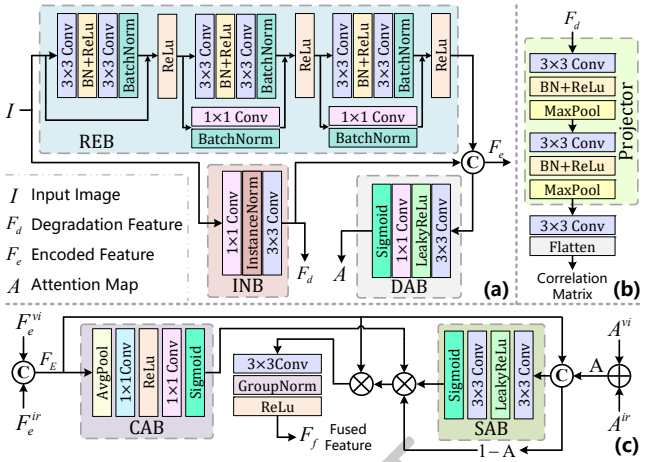

Figure 5: (a) The proposed REB, INB and DAB. (b) The proposed projector in DSDM. (c) The proposed DGFM, including a channel attention block (CAB) and a spatial attention block (SAB).

3.2 Fusion Pipline

Image Encoder

The proposed network employs separate infrared and visible encoders with identical architectures but non-shared weights, as shown in Figure 5. Due to the statistical characteristics of different degradations may exhibit significant variations, which makes optimization difficult, an instance normalization module is utilized to perform rough alignment of low-level statistical features for input images. Meanwhile, since the normalization process inevitably causes information loss that may lead to information loss, a parallel feature extraction branch is constructed based on residual blocks [He *et al.*, 2016] to ensure complete transmission of source image information to the fusion module. The features extracted from two paths are concatenated and sent to the fusion module.

Specifically, the instance normalization block (INB) is performed as follows:

$$IN(x) = \gamma \frac{x - \mu(x)}{\sigma(x)} + \beta, \quad (3)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation of each channel, γ and β are learnable parameters. The output of INB is degradation feature F_d.

The residual feature extraction branch (REB) comprises three sequentially connected residual blocks. The encoded feature F_e are the concatenation of F_d and output features of the REB. The degradation-aware attention map A is calculated from F_e.

Degradation-Aware Gated Fusion Module

Furthermore, to adaptively adjust fusion strategies according to decoupled degradation-type-specific feature, we propose a degradation-aware gated fusion module that adaptively adjusts the fusion strategy in both spatial and channel dimensions based on the previously computed degradation-aware attention maps, as shown in Figure 5.

Specifically, it first compute degradation attention maps from the encoded feature in DAB. And we concatenate the

Figure 6: Qualitative comparison results in practical scenario.

encoded features F_e of both modalities along the channel dimension and sum their corresponding degradation-aware attention maps A to obtain F_E and A. Subsequently, channel attention and spatial attention are computed from F_E, A and the inversion of A. Then element-wise multiplication is preformed on these components to generate the gating weights G for F_E. Finally, the fused feature F_f is obtained by element-wise multiplication between G and F_E.

Fusion Decoder and Loss

The fusion decoder D consists of upsampling and downsampling operations with a residual block interleaved between them, which transforms the fused feature representation F_f into the output fused image I_f. The loss functions are computed between I_f and high-quality source images I_{vi}^{gt} and I_{ir}^{gt}.

The max intensity loss L_{max} calculates the intensity discrepancy between fused image and the supremum of corresponding source images. It is expressed as:

$$L_{max} = \frac{1}{HW}\left\| I_f - max\left(I_{vi}^{gt}, I_{ir}^{gt}\right)\right\|_1, \tag{4}$$

where H and W represent the height and width of the image,

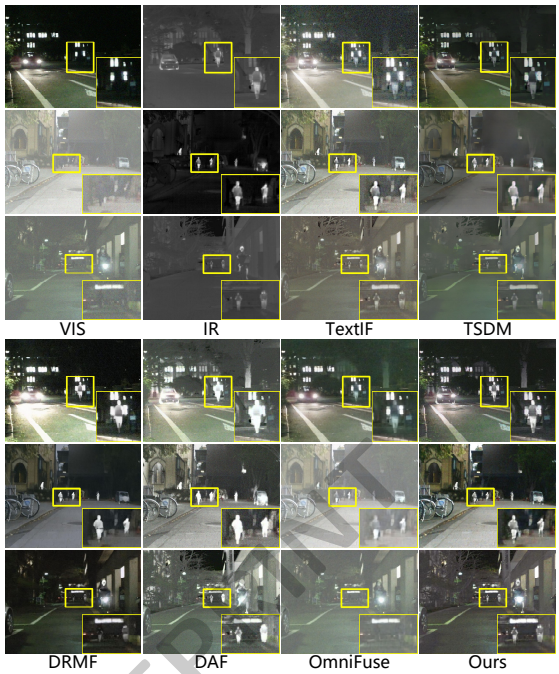

Figure 7: Qualitative comparison results under visible composite degradation. The exhibited samples from top to bottom are Degradation I to III, respectively.

$max\left(\cdot\right)$ denotes the element-wise maximum selection.

The structural similarity (SSIM) [Wang *et al.*, 2004] loss function quantifies structural similarity between the fused image and source images by evaluating luminance, contrast, and structural patterns. It is defined as:

$$L_{SSIM} = \left(1 - ssim\left(I_f, I_{vi}^{gt}\right)\right) + \left(1 - ssim\left(I_f, I_{ir}^{gt}\right)\right). \tag{5}$$

The gradient loss computes the loss of higher-intensity edges between the source images, which can be expressed as follows, where ∇ denotes the Sobel operator:

$$L_{grad} = \frac{1}{HW}\left\| \nabla I_f - max\left(\nabla I_{vi}^{gt}, \nabla I_{ir}^{gt}\right)\right\|_1. \tag{6}$$

The overall loss function consists of degradation-type specificity decoupling loss and fusion loss:

$$L_{total} = \alpha L_{SPE} + \beta L_{max} + \gamma L_{SSIM} + \delta L_{grad}, \tag{7}$$

where α, β, γ and δ are respectively set to 1e-5, 4, 1 and 1 for the balance between loss terms.

4 Experiment

4.1 Experimental Details

The proposed SpeciFuse needs high-quality infrared-visible image pairs as benchmarks due to its capability of learning degradation-type specificity between different individual degradations in same scene. For this purpose, we selected 718 precisely aligned infrared-visible image pairs from the EMS [Yi *et al.*, 2024] dataset to construct the training set. And we randomly extracted 128×128 co-located patches from each source image pair as network inputs. The entire network

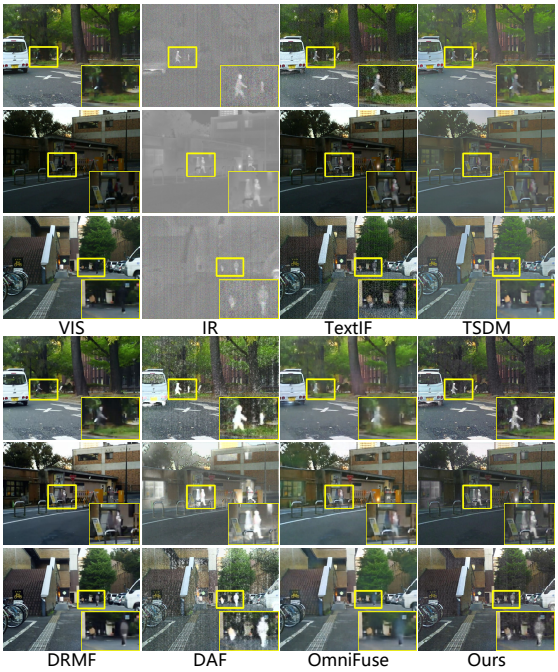

Figure 8: Qualitative comparison results under infrared composite degradation. The exhibited samples from top to bottom are Degradation IV to VI, respectively.

Method	M3FD			LLVIP			MSRS		
	PSNR↑	CC↑	SF↑	PSNR↑	CC↑	SF↑	PSNR↑	CC↑	SF↑
YDTR	62.120	0.602	9.700	62.743	0.663	15.901	61.959	0.629	7.388
CDDFuse	62.741	0.599	14.510	62.211	0.689	14.681	62.051	0.600	11.556
LRRNet	62.115	0.581	10.352	61.024	0.669	15.800	62.720	0.515	8.555
PIAFusion	61.952	0.595	14.858	62.773	0.682	15.266	62.098	0.600	12.130
TextIF	62.178	0.612	14.754	62.503	0.680	19.043	62.765	0.598	11.893
TSDM	62.890	0.608	12.071	61.583	0.703	14.881	61.767	0.646	9.399
DRMF	60.544	0.509	15.117	60.140	0.582	18.415	59.699	0.517	12.702
DAF	61.033	0.602	15.096	58.529	0.707	18.112	57.905	0.628	13.817
OmniFuse	62.240	0.566	11.679	60.450	0.663	11.617	60.823	0.435	12.563
Ours	**62.918**	**0.601**	**16.915**	**62.822**	**0.706**	**19.777**	**62.939**	**0.639**	**12.987**

Table 1: Quantitative comparison of metrics in practical scenarios. The best and second-best are marked in red and blue, respectively.

was implemented in PyTorch and trained on an NVIDIA 3090 GPU with empirically determined hyperparameters: the infrared and visible encoders were optimized with a learning rate of 1e-5, while the gated fusion module and fusion decoder employed a learning rate of 1e-4. All components were jointly trained for 60 epochs using the AdamW [Loshchilov and Hutter, 2017] optimizer with a batch size of 4.

4.2 Model Comparisons

To demonstrate the effectiveness of the proposed method, we conducted comprehensive comparisons with nine state-of-the-art (SOTA) fusion approaches. These include four fusion methods for the normal scenario (YDTR [Tang *et al.*, 2022b], CDDFuse [Zhao *et al.*, 2023], LRRNet [Li *et al.*, 2023] and PIAFusion [Tang *et al.*, 2022a]), and five degradation-resistant fusion algorithms (TextIF [Yi *et al.*, 2024], TSDM [Xue *et al.*, 2024], DRMF [Tang *et al.*, 2024], DAF [Wang *et al.*, 2025], and OmniFuse [Zhang *et al.*, 2025]). For evaluating practical fusion performance, we employ benchmark datasets including M3FD [Liu *et al.*, 2022], LLVIP [Jia *et al.*, 2021], and MSRS [Tang *et al.*, 2022a]. Based EMS dataset, we synthesize low-quality images for assessing fusion capabilities under composite degradations.

The quantitative metrics employed in this study include Peak Signal-to-Noise Ratio (PSNR), Cross-Correlation coefficient (CC), Spatial Frequency (SF) [Eskicioglu and Fisher, 2002], Mean Square Error (MSE), Spectral Correlation Distortion (SCD) [Aslantas and Bendes, 2015], and Quality Assessment of image Fusion based on edge information (Qabf) [Xydeas and Petrovic, 2000].

Image Fusion in Practical Scenario

The qualitative results under practical scenario is shown in Figure 6. Specifically, YDTR, LRRNet, and TSDM produce insufficient brightness in target regions and background contrast. CDDFuse, PIAFusion, and TextIF maintain basic thermal intensity and details but lack target prominence and color fidelity, with suboptimal brightness. OmniFuse yields overly smooth results, losing texture sharpness. Although DRMF and DAF achieve higher brightness, both lose intrinsic contrast and internal detail within thermal targets. Additionally, DRMF weakens thermal targets, while DAF introduces border expansion artifacts. In contrast, SpeciFuse achieves the most balanced and informative fusion.

Quantitative results are presented in Table 1. SpeciFuse achieves strong overall performance across PSNR, CC, and SF, reflecting its superiority in preserving structural detail, maintaining source consistency, and enhancing visual clarity

Image Fusion in Composite Degradation Scenario

To further assess fusion performance under complex degradations, we constructed a composite degradation test set based on the EMS dataset, with separate sets for visible and infrared modalities. For visible images, the set includes three composite degradations: low-light + noise (**Degradation I:**), low-light + haze (**II**), and low-light + random noise + haze (**III**). For infrared images, it includes three composite degradations: low-contrast + random noise + stripe noise (**IV**), blur + low-contrast (**V**), and low-contrast + random noise + stripe noise + blur (**VI**). For TextIF, we used "low-light" and "low-contrast" text prompts corresponding to the visible and infrared degradations, respectively.

Qualitative results in composite degradations are shown in Figures 7 and 8. TextIF reduces degradation impact via text guidance but is limited to predefined categories and declines on unseen degradations. TSDM denoises well but underperforms in brightness and contrast restoration. DRMF generates generally good quality but fails under some conditions, and it over-relies on visible information, covering the infrared information. DAF handles unknown degradations but amplifies noise and causes over-enhancement, and it loses internal contrast of objects. OmniFuse suffers from overall blurring and weak thermal objects. In contrast, SpeciFuse preserves both object saliency and scene details more comprehensively.

Quantitative results under six degradation scenarios are summarized in Table 2, where our method achieves strong overall performance. The superior PSNR and lower MSE reflect high pixel-level fidelity and effective detail recovery

Method	Degradation I				Degradation II				Degradation III				Degradation IV				Degradation V				Degradation VI			
	PSNR↑	MSE↓	SCD↑	Qabf↑	PSNR↑	MSE↓	SCD↑	Qabf↑	PSNR↑	MSE↓	SCD↑	Qabf↑	PSNR↑	MSE↓	SCD↑	Qabf↑	PSNR↑	MSE↓	SCD↑	Qabf↑	PSNR↑	MSE↓	SCD↑	Qabf↑
TextIF	63.748	0.031	1.172	0.328	58.764	0.093	1.084	0.387	63.267	0.045	0.649	0.307	60.571	0.056	1.210	0.378	61.396	0.046	1.387	0.567	61.630	0.060	1.140	0.368
TSDM	64.199	0.026	1.197	0.338	62.430	0.039	1.222	0.221	64.541	0.025	0.925	0.252	60.826	0.058	1.156	0.461	60.791	0.058	1.160	0.488	60.812	0.057	1.150	0.461
DRMF	62.277	0.041	1.082	0.285	62.678	0.039	1.171	0.381	64.069	0.028	0.772	0.250	59.833	0.073	1.063	0.552	59.950	0.071	1.126	0.551	59.837	0.073	1.047	0.482
DAF	61.129	0.052	1.485	0.305	59.148	0.083	1.487	0.300	61.081	0.053	1.253	0.272	58.917	0.088	1.426	0.244	58.976	0.088	1.570	0.464	58.835	0.090	1.401	0.238
OmniFuse	65.446	0.021	1.221	0.326	56.759	0.144	0.857	0.172	60.340	0.068	0.471	0.164	59.612	0.078	1.082	0.408	60.231	0.065	1.246	0.452	59.625	0.078	1.067	0.408
Ours	**65.609**	**0.020**	**1.235**	**0.331**	**61.954**	**0.056**	**1.492**	**0.411**	**65.755**	**0.020**	**1.157**	**0.296**	**60.883**	**0.057**	**1.441**	**0.472**	**61.030**	**0.055**	**1.496**	**0.553**	**60.832**	**0.058**	**1.423**	**0.461**

Table 2: Quantitative comparison of metrics under composite degradation. The best and second-best are marked in red and blue, respectively.

Method	M3FD			MSRS			LLVIP		
	PSNR↑	CC↑	SF↑	PSNR↑	CC↑	SF↑	PSNR↑	CC↑	SF↑
w/o DSDM	62.162	0.540	14.593	62.647	0.662	14.819	60.774	0.637	11.668
w/o DGFM	62.750	0.500	15.013	60.691	0.649	16.327	60.235	0.577	11.977
w/o L_{max}	62.860	0.540	16.526	62.198	0.651	19.592	61.075	0.603	11.859
w/o L_{SSIM}	62.694	0.530	16.579	61.973	0.655	20.387	60.534	0.585	12.224
w/o L_{grad}	62.773	0.542	15.534	62.415	0.664	17.362	60.717	0.604	11.747
Ours	**62.918**	**0.551**	**16.915**	**62.422**	**0.666**	**19.777**	**62.939**	**0.639**	**12.987**

Table 3: Quantitative comparison of ablation study. Higher values of PSNR, CC, and SF indicate superior fusion quality.

		$\lambda=10^{-2}$	$\lambda=10^{-3}$	$\alpha=10^{-4}$	$\alpha=10^{-6}$	$\beta=2$	$\beta=8$	$\gamma=0.5$	$\gamma=2$	$\delta=0.5$	$\delta=2$	**ours**
Deg.I	PSNR	65.34	65.46	65.14	65.18	65.14	65.12	65.46	65.43	64.08	65.53	**65.61**
	MSE	0.023	0.022	0.024	0.023	0.023	0.023	0.022	0.023	0.028	0.022	**0.020**
	SCD	1.222	1.099	1.276	1.211	1.162	1.174	1.308	1.240	1.072	1.213	**1.235**
	Qabf	0.323	0.305	0.322	0.322	0.313	0.301	0.307	0.329	0.209	0.310	**0.331**
Deg.II	PSNR	60.65	61.18	60.07	60.85	61.50	61.45	61.15	60.37	61.52	61.06	**61.95**
	MSE	0.067	0.061	0.076	0.064	0.057	0.058	0.061	0.071	0.055	0.062	**0.056**
	SCD	1.468	1.457	1.549	1.572	1.601	1.490	1.540	1.480	1.375	1.490	**1.492**
	Qabf	0.390	0.373	0.385	0.388	0.389	0.390	0.385	0.391	0.352	0.369	**0.411**
Deg.III	PSNR	64.84	65.01	64.65	64.57	65.02	64.88	65.07	64.91	63.76	65.11	**65.76**
	MSE	0.021	0.020	0.022	0.022	0.020	0.021	0.020	0.021	0.025	0.020	**0.020**
	SCD	1.156	1.050	1.117	1.108	1.118	1.149	1.190	1.172	1.011	1.150	**1.157**
	Qabf	0.288	0.274	0.288	0.287	0.283	0.277	0.281	0.292	0.205	0.278	**0.296**
Deg.IV	PSNR	60.79	61.08	60.61	61.03	60.54	60.77	60.46	59.62	58.57	60.18	**60.88**
	MSE	0.059	0.058	0.065	0.059	0.059	0.061	0.064	0.064	0.081	0.057	**0.057**
	SCD	1.317	1.273	1.400	1.389	1.416	1.389	1.401	1.407	1.175	1.405	**1.441**
	Qabf	0.459	0.441	0.459	0.454	0.404	0.385	0.385	0.416	0.308	0.386	**0.472**
Deg.V	PSNR	60.68	60.97	60.35	60.68	61.28	60.52	60.13	60.41	59.87	60.82	**61.03**
	MSE	0.056	0.058	0.058	0.053	0.056	0.058	0.053	0.062	0.071	0.056	**0.055**
	SCD	1.443	1.437	1.498	1.487	1.537	1.414	1.444	1.437	1.346	1.475	**1.496**
	Qabf	0.521	0.509	0.527	0.519	0.527	0.501	0.505	0.536	0.537	0.505	**0.553**
Deg.VI	PSNR	59.97	60.05	60.56	60.01	60.92	60.75	60.43	60.59	59.53	60.16	**60.83**
	MSE	0.062	0.062	0.068	0.062	0.062	0.054	0.057	0.067	0.084	0.060	**0.058**
	SCD	1.393	1.344	1.421	1.414	1.432	1.401	1.416	1.421	1.199	1.416	**1.423**
	Qabf	0.410	0.392	0.411	0.406	0.415	0.396	0.397	0.428	0.314	0.398	**0.461**

Table 4: Quantitative results of hyperparameter sensitivity analysis.

across degradations. Outstanding SCD and Qabf scores indicate a superior balance between thermal contrast retention and visible texture preservation, along with strong edge and directional feature maintenance. These results align with qualitative observations and demonstrate that SpeciFuse robustly recovers both structural and textural information under composite degradations.

Futhermore, SpeciFuse achieves a highly competitive balance between efficiency and performance, with 1.170M trainable parameters and an average runtime of 0.093 seconds per image (640×480 px) on an NVIDIA RTX3090 GPU. In comparison, the corresponding performance for other degradation-resistant methods are: TextIF (215.110M, 0.387s), TSDM (0.635M, 0.965s), DRMF (170.90M, 9.782s), DAF (0.278M, 0.085s), and OmniFuse (173.34M, 5.372s).

4.3 Ablation Study

In order to validate DSDM, DGFM and loss functions, we perform five set of ablation experiments on them respectively. Specifically, "w/o DSDM" represents the variant that remove degradation-type specificity decoupling module and decoupling loss in training stage. In "w/o DGFM", the visible and infrared encoded features were simply concatenated before being fed into the decoder. "w/o L_{SSIM}", "w/o L_{max}" and "w/o L_{grad}" respectively denote the variant that remove the corresponding loss function in training stage. Figure 9 shows ablation results under a practical scene. Removing the DSDM and DGFM leads to failure in restoring color and details from darkness, reduced target contrast, and blurred details. Omitting the intensity loss causes severe color distortion. Removing the SSIM loss slightly degrades brightness and details, while dropping the gradient loss blurs textures. These findings confirm that both modules are essential for adaptive restoration and preservation, and each loss is critical for visual fidelity. Furthermore, we conducted hyperparameter sensitivity analysis on λ, α, β, γ and δ. Quantitative results in Table 3 and 4 show the full model significantly outperforms all ablated variants, verifying that every component is indispensable for optimal performance.

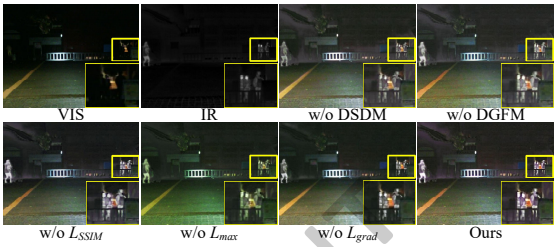

Figure 9: Qualitative comparison results of ablation study.

5 Conclusion

This study proposes SpeciFuse, a robust infrared-visible image fusion network that learns the specificity of individual degradations via a siamese architecture, thus modeling and addressing the conflicts arising from composite degradations in the fusion process. In SpeciFuse, the degradation-type specificity decoupling module is designed between the encoders of two pipelines to decorrelate the feature representations of different degradation types, thereby reducing their mutual interference and preserving structural information in the feature encoding. Furthermore, a degradation-aware gated fusion module is proposed to adaptively learn degradation-conditional fusion strategies based on the decoupled features. This module dynamically modulates fusion process via channel-wise and spatial-wise gating mechanisms, significantly enhancing the fusion robustness and quality across diverse conditions. Remarkably, although trained on single-degradation conditions, comprehensive experiments demonstrate the superiority of SpeciFuse in handling diverse composite degradations.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No.62401411. This work is also supported by Natural Science Foundation of Hubei Province under Grant No.2026AFB616.

References

[Aslantas and Bendes, 2015] Veysel Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of electronics and communications*, 69(12):1890–1896, 2015.

[Chen *et al.*, 2024] Jun Chen, Liling Yang, Wei Liu, Xin Tian, and Jiayi Ma. Lenfusion: a joint low-light enhancement and fusion network for nighttime infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024.

[Chen *et al.*, 2025] Liyuan Chen, Wei Liu, Hua Wang, Sang-Woon Jeon, Yunliang Jiang, and Zhonglong Zheng. Consistency-guided adaptive alternating training for semi-supervised salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[Eskicioglu and Fisher, 2002] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 2002.

[Guo *et al.*, 2024] Jinxin Guo, Weida Zhan, Yichun Jiang, Wei Ge, Yu Chen, Xiaoyu Xu, Jin Li, and Yanyan Liu. Mfhod: Multi-modal image fusion method based on the higher-order degradation model. *Expert Systems with Applications*, 249:123731, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[Jia *et al.*, 2021] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.

[Lei *et al.*, 2024] Xinyu Lei, Longjun Liu, Puhang Jia, Haoteng Li, and Haonan Zhang. Low-light infrared and visible image fusion with imbalanced thermal radiation distribution. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[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.*, 2022] Xuan Li, Yuhang Xu, Lei Ma, Zhenghua Huang, and Haiwen Yuan. Progressive attention-based feature recovery with scribble supervision for saliency detection in optical remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.

[Li *et al.*, 2023] 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 Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023.

[Li *et al.*, 2025a] Xuan Li, Rongfu Chen, Jie Wang, Weiwei Chen, Huabing Zhou, and Jiayi Ma. Caspfuse: An infrared and visible image fusion method based on dual-cycle crosswise awareness and global structure-tensor preservation. *IEEE Transactions on Instrumentation and Measurement*, 74:1–15, 2025.

[Li *et al.*, 2025b] Xuan Li, Cheng Zhang, Jie Wang, Rongfu Chen, and Li Cheng. Bidirectional feedback network for high-level task-directed infrared and visible image fusion. *Infrared Physics & Technology*, 147:105751, 2025.

[Li *et al.*, 2025c] Xuan Li, Guomin Zhang, Weiwei Chen, Li Cheng, Yining Xie, and Jiayi Ma. An infrared and visible image fusion method based on semantic-sensitive mask selection and bidirectional-collaboration region fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(6):5518–5532, 2025.

[Liu *et al.*, 2022] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

[Liu *et al.*, 2024] Xiaowen Liu, Hongtao Huo, Jing Li, Shan Pang, and Bowen Zheng. A semantic-driven coupled network for infrared and visible image fusion. *Information Fusion*, 108:102352, 2024.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[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.*, 2020] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

[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.

[Tang *et al.*, 2022a] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022.

[Tang *et al.*, 2022b] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 25:5413–5428, 2022.

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

[Tang *et al.*, 2024] Linfeng Tang, Yuxin Deng, Xunpeng Yi, Qinglong Yan, Yixuan Yuan, and Jiayi Ma. Drmf: Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8546–8555, 2024.

[Tang *et al.*, 2025a] Linfeng Tang, Chunyu Li, and Jiayi Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025.

[Tang *et al.*, 2025b] Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion framework with language-vision degradation prompts. *Advances in Neural Information Processing Systems*, 2025.

[Tang *et al.*, 2025c] Linfeng Tang, Qinglong Yan, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. C2rf: Bridging multimodal image registration and fusion via commonality mining and contrastive learning. *International Journal of Computer Vision*, 133:5262–5280, 2025.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wang *et al.*, 2025] 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*, 118:102931, 2025.

[Xu *et al.*, 2021] Han Xu, Hao Zhang, and Jiayi Ma. Classification saliency-based rule for visible and infrared image fusion. *IEEE Transactions on Computational Imaging*, 7:824–836, 2021.

[Xue *et al.*, 2024] Weimin Xue, Yisha Liu, Fei Wang, Guojian He, and Yan Zhuang. A novel teacher–student framework with degradation model for infrared–visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 73:1–12, 2024.

[Xydeas and Petrovic, 2000] Costas S Xydeas and Vladimir Petrovic. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000.

[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.

[Zbontar *et al.*, 2021] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the International Conference on Machine Learning*, pages 12310–12320, 2021.

[Zhang and Demiris, 2023] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023.

[Zhang *et al.*, 2020] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12797–12804, 2020.

[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*, 47(9):7577–7595, 2025.

[Zhao *et al.*, 2023] 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.*, 2024] Shenlu Zhao, Jingyi Li, and Qiang Zhang. C 4 net: Excavating cross-modal context-and content-complementarity for rgb-t semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.