MdaIF: Robust One-Stop Multi-Degradation-Aware Image Fusion with Language-Driven Semantics

Jing Li[1,2], **Yifan Wang**[3], **Jiafeng Yan**[3], **Renlong Zhang**[3], **Bin Yang**[4*]

[1]Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China

[2]Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, Ministry of Natural Resources, East China Normal University, Shanghai 200241, China

[3]School of Information, Central University of Finance and Economics, Beijing 102206, China

[4]School of Artificial Intelligence and Robotics, Hunan University, Changsha 410082, China

jingli@geoai.ecnu.edu.cn, {2024212461, yan.jiafeng, 2023312325}@email.cufe.edu.cn, binyang@hnu.edu.cn

Abstract

Infrared and visible image fusion aims to integrate complementary multi-modal information into a single fused result. However, existing methods 1) fail to account for the degradation visible images under adverse weather conditions, thereby compromising fusion performance; and 2) rely on fixed network architectures, limiting their adaptability to diverse degradation scenarios. To address these issues, we propose a one-stop degradation-aware image fusion framework for multi-degradation scenarios driven by a large language model (MdaIF). Given the distinct scattering characteristics of different degradation scenarios (e.g., haze, rain, and snow) in atmospheric transmission, a mixture-of-experts (MoE) system is introduced to tackle image fusion across multiple degradation scenarios. To adaptively extract diverse weather-aware degradation knowledge and scene feature representations, collectively referred to as the semantic prior, we employ a pre-trained vision-language model (VLM) in our framework. Guided by the semantic prior, we propose degradation-aware channel attention module (DCAM), which employ degradation prototype decomposition to facilitate multi-modal feature interaction in channel domain. In addition, to achieve effective expert routing, the semantic prior and channel-domain modulated features are utilized to guide the MoE, enabling robust image fusion in complex degradation scenarios. Extensive experiments validate the effectiveness of our MdaIF, demonstrating superior performance over SOTA methods.

Code — https://github.com/doudou845133/MdaIF

Extended version — https://doi.org/10.48550/arXiv.2511.12525

Introduction

Infrared and visible image fusion (IVF) aims to integrate the complementary advantages of multi-modal sensors (Zhang et al. 2021; Liu et al. 2024). Infrared sensor captures thermal radiation from objects, presenting thermodynamic characteristics via pixel intensity distributions but lacking fine-grained texture details. In contrast, visible sensor relies on surface reflectance to provide rich texture information. IVF

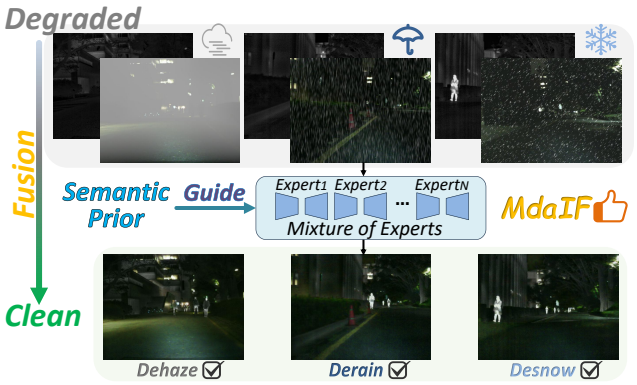

Figure 1: Our fusion results under haze, rain, and snow conditions, producing clean outputs from degraded inputs.

significantly enhances scene perception and offers more robust inputs for downstream vision tasks such as assisted driving (Bao et al. 2023) and intelligent surveillance (Zhang et al. 2018). However, in the imaging process, due to the wavelength difference between infrared and visible light, adverse weather conditions such as haze, rain, and snow cause significant scattering of visible light, whereas infrared imaging remains relatively unaffected, especially over short distances (Judd, Thornton, and Richards 2019; Velázquez et al. 2022). As a result, texture details in visible images are poorly preserved, thereby impairing the performance of IVF and further affecting high-level vision tasks. Therefore, *accounting for adverse weather degradations in IVF is crucial to enhancing its generalization and applications*.

To address the above issue, with the advancement of image restoration (Yang et al. 2024b; Conde, Geigle, and Timofte 2024) and multi-modal image fusion (Wu et al. 2025; Zhao et al. 2024) technologies, a straightforward approach to infrared and degraded visible image fusion (IdVF) is to cascade image restoration and multi-modal fusion models, as Fig. 2 (a), however, such a sequential design often leads to suboptimal feature alignment and error accumulation across tasks. In addition, we cascade SOTA image restoration and fusion models for IdVF, and experiments show this strat-

*Corresponding author

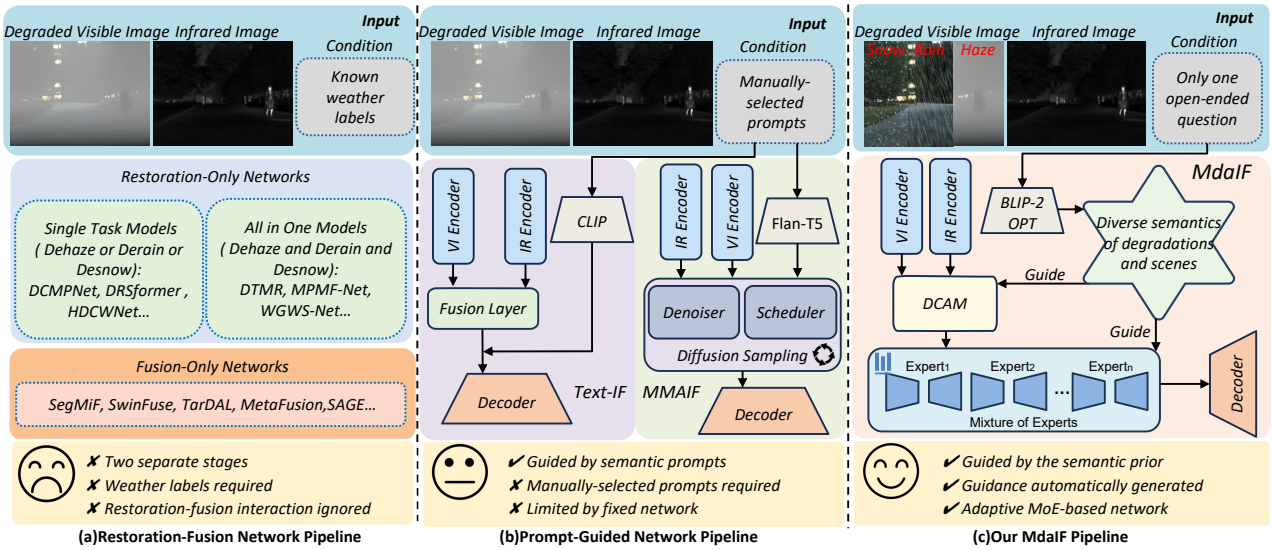

Figure 2: Comparisons with previous pipelines.

egy is insufficient to address the above issue, as detailed in the experimental section. In contrast, Fig. 1 shows that our method yields superior fusion results on degraded visible and infrared inputs.

IdVF has been gaining momentum as a prominent research focus, several methods have begun to incorporate VLMs or large language models (LLMs) to enhance performance in degraded image fusion scenarios. For example, Text-IF (Yi et al. 2024) leverages prompt-guided mechanisms built upon CLIP (Radford et al. 2021), a typical VLM, to facilitate image fusion in degradation scenarios like low illumination and overexposure, yet it overlooks the adverse weather-induced degradations that significantly affect fusion performance. In addition, MMAIF (Cao et al. 2025) combines diffusion model with Flan-T5 (Chung et al. 2024), a representative LLM, to perform image fusion under diverse degradations. However, both MMAIF and Text-IF rely on fixed prompts derived from ground-truth degradation types and employ a single unified network for all conditions, which limits the performance of IdVF, as depicted in Fig. 2 (b). Accordingly, we question: *can IdVF under diverse adverse weather conditions be adaptively achieved by leveraging multiple networks to generate clean fused image, without requiring relying on ground-truth degradation types?*

Therefore, we propose a one-stop degradation-aware image fusion framework for multi-degradation scenarios driven by VLM. **The motivations are shown as follow**:

1) *The distinct atmospheric scattering characteristics of haze, rain, and snow hinder the ability of a fixed network to generalize across different degradation scenarios.* Specifically, the differences in particle characteristics—micron-sized droplets in haze, millimeter-scale raindrops in rain, and ice crystals in snow—lead to fundamentally different atmospheric scattering models. Fixed network architectures exhibit limited capacity in capturing the heterogeneous degradation patterns associated with various adverse weather conditions. For example, the transmission map, which is essential for haze removal (Song et al. 2023), becomes ineffective when applied to deraining scenarios (Fu et al. 2017).

To this end, we propose a MoE-based framework, where a set of specialized experts are tailored to effectively solve the IdVF task under multiple degradation conditions. To mitigate the performance degradation caused by imbalanced expert selection—where one expert is overloaded with multiple tasks while others remain inactive—we propose a semantic prior-guided expert routing strategy. The proposed strategy employs the semantic prior derived from the VLM to interact with the modulated features, establishing a task-specific expert routing mechanism. This enables the model to adaptively select and compose appropriate experts based on prior knowledge, thereby effectively addressing IdVF tasks under diverse degradation scenarios.

2) *Existing IdVF methods rely on ground-truth degradation types as prompts, which limits their flexibility and constrains their applicability.* Specifically, relying on fixed prompts derived from ground-truth degradation types merely exploits the text-visual alignment capabilities of these models, without leveraging their potential to understand complex degradation scenarios, thereby limiting the flexibility of IdVF under diverse degradation conditions.

Therefore, we introduce the VLM into our method, which fully leverages the scene understanding capabilities—demonstrated effective in restoration tasks (Yang et al. 2024b)—to enable the model to adaptively perceive degradation types, thereby guiding MoE to select appropriate task-specific experts. Rather than using the VLM solely as a classifier for degradation types, we further exploit its semantic understanding of the scene to enhance feature interaction and representation within the DCAM module. Our contributions can be summarized as follows:

- Our method breaks the conventional reliance on sequen-

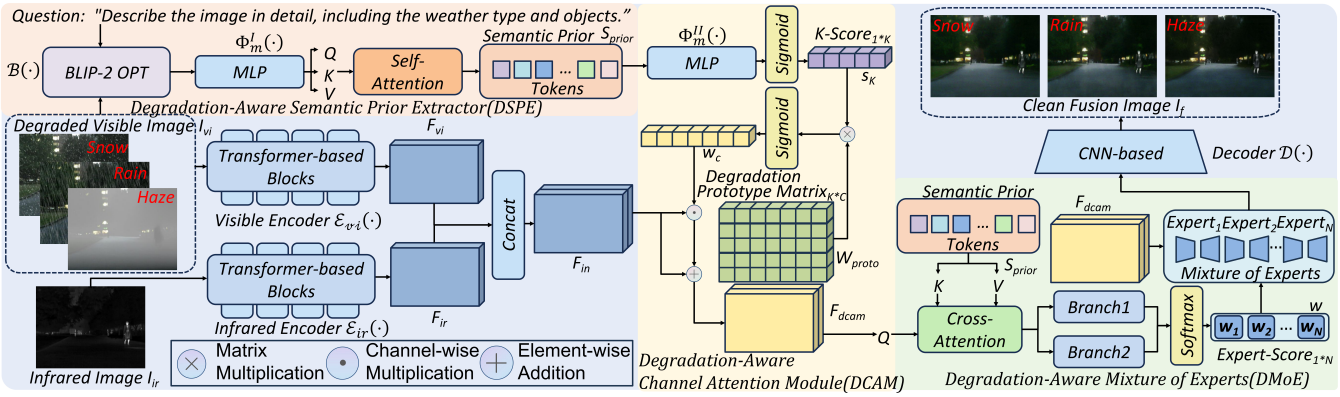

Figure 3: Overview of the proposed network architecture.

tial subtask processing by introducing joint optimization for degraded multi-modal fusion, thus alleviating feature misalignment and error accumulation.

- To break the dependency on ground-truth degradation types and improve the model's adaptability and applicability, we propose a VLM-based adaptive degradation-aware framework for multi-degradation IdVF.
- To address the limitations of fixed networks in modeling heterogeneous atmospheric degradations, we propose a MoE-based degradation-aware image fusion framework, which is guided by the semantic prior of the VLM and feature representations strengthened by DCAM.

Related Work

General Image Fusion

IVF has emerged as a critical subfield of computer vision and has received extensive attention in recent years (Ma, Ma, and Li 2019; Zhang and Demiris 2023). With the advancement of algorithms, IVF has evolved from traditional manually designed fusion rules to end-to-end deep learning-based frameworks, which includes CNN- and GAN-based local feature fusion methods (Ren et al. 2018; Li et al. 2019), Transformer-based global feature fusion methods (Yang et al. 2024a), hybrid models that integrate local and global features (Chen et al. 2023a), and diffusion-based denoising fusion frameworks (Zhao et al. 2023b). In terms of fusion objectives, IVF has evolved from visually-oriented methods to task-driven fusion strategies designed to support downstream applications such as semantic segmentation (Liu et al. 2023b) and object detection (Jiang et al. 2024). With respect to fusion conditions, IVF has developed from the fusion of well-registered source images to that of unregistered inputs (Li et al. 2025). However, existing methods rarely consider or simultaneously address multi-type degradation caused by diverse adverse weather scenarios.

Prompt-Guided Image Fusion and Degradation-Awareness

Driven by the rapid advancement of VLMs and LLMs, recent studies have integrated these models into IVF frameworks to enhance adaptability and flexibility. On one hand, such models offer improved capabilities for adaptive fusion by enabling fine-grained control over specific regions of interest. For example, TeRF (Wang et al. 2024) employs text-driven and region-aware mechanisms to achieve semantically guided IVF. On the other hand, they have been utilized to facilitate degradation-aware IVF. For instance, TexIF leverages prompt-based guidance to perceive specific types of degradations and perform corresponding IdVF. Additionally, OmniFuse (Zhang et al. 2025) and MMAIF integrate semantic prompts with diffusion model to address degradation-aware fusion tasks, and OmniFuse supports prompt-based region-specific enhancement to improve the flexible of IVF.

However, they overlook the heterogeneous degradation factors caused by diverse adverse weather conditions. Furthermore, all the aforementioned methods rely on ground-truth degradation labels and fixed networks, without considering the distinct transmission patterns and mathematical formulations associated with different types of degradation. Therefore, we propose a VLM-based adaptive degradation-aware framework for multi-degradation IdVF.

Method

As illustrated in Fig. 3, the MdaIF framework is fundamentally guided by semantic priors. In the following sections, we present a comprehensive analysis of our method from three key perspectives: problem formulation, network architecture, and loss functions.

Problem Formulation

General IVF methods typically adopt a fixed network θ_n and implicitly assume a clean visible image $\tilde{I_{vi}}$, without considering the effects of severe degradation. The network is designed to learn a predefined fusion function $\mathcal{F}_{if}(\cdot)$ for generating the fusion result. The fusion process can be formulated as:

$$I_f = \mathcal{F}_{if}(\tilde{I_{vi}}, I_{ir}; \theta_n). \quad (1)$$

However, in adverse weather, visible images often suffer severe degradation that fixed networks cannot handle effectively. Thus, we design an adaptive MoE-based IdVF network guided by semantic priors, which enable degradation-

aware fusion without relying on ground-truth degradation labels. The process is defined as:

$$I_f = \mathcal{F}_{mdaif}(I_{vi}, I_{ir}, S_{prior}; \theta_{moe}), \quad (2)$$

where I_{vi} denotes the degraded visible image, S_{prior} represents the semantic prior derived from the VLM, $\mathcal{F}_{mdaif}(\cdot)$ is the multi-degradation-aware fusion function, and θ_{moe} indicates the parameters of the MoE-based network.

Network Structure

Encoder In our fusion pipeline, the degraded visible image $I_{vi} \in \mathbb{R}^{H\times W\times 3}$ and the infrared image $I_{ir} \in \mathbb{R}^{H\times W\times 1}$ are independently encoded by encoders $\mathcal{E}_{vi}$ and $\mathcal{E}_{ir}$, which are configured as transformer-based structures following SegFormer (Xie et al. 2021). Each encoder comprises 4 transformer blocks. The encoded features F_{vi} and F_{ir} are concatenated along the channel dimension and subsequently fed into DCAM, which is formulated as:

$$F_{vi} = \mathcal{E}_{vi}(I_{vi}), F_{ir} = \mathcal{E}_{ir}(I_{ir}), F_{in} = \mathrm{Cat}(F_{vi}, F_{ir}), \quad (3)$$

where the operator $\mathrm{Cat}(\cdot)$ denotes channel-wise concatenation, and $F_{in} \in \mathbb{R}^{H\times W\times C}$, with C representing the number of channels.

Degradation-Aware Semantic Prior Extractor (DSPE) We employ BLIP-2 OPT 2.7B (Li et al. 2023), a pre-trained VLM, to extract semantic prior from the degraded visible image. Specifically, we adopt a visual question answering (VQA) paradigm, where the degraded visible image I_{vi} and an open-ended question prompt $\mathcal{P}_q$ are jointly fed into the VLM to generate degradation-aware original semantic prior, denoted as $S_{org} \in \mathbb{R}^{S\times C_{org}}$, where S represents the sequence length and C_{org} denotes the original token dimension. We formulate it as:

$$S_{org} = \mathcal{L}(\mathcal{B}(I_{vi}, \mathcal{P}_q)), \quad (4)$$

where $\mathcal{B}(\cdot)$ denotes the BLIP-2 OPT model, and $\mathcal{L}(\cdot)$ represents the operator used to extract the features from its last hidden layer.

To better align the original semantic prior with the feature space of our framework, we refine it through an MLP $\Phi_m^I(\cdot)$ to reduce the embedding dimension, producing the compressed representation $S_{embed} \in \mathbb{R}^{S\times C}$:

$$S_{embed} = \mathcal{N}_{layer}(\Phi_m^I(S_{org})), \quad (5)$$

where $\mathcal{N}_{layer}(\cdot)$ denotes layer normalization, which promotes stable training and faster convergence. Here, C denotes the operational feature dimension in our framework.

To emphasize the most salient tokens, we apply a self-attention mechanism to adaptively reweight their importance in the semantic prior. This adaptive refinement produces a well-aligned and semantically enriched prior, denoted as S_{prior}, which provides a robust and informative guidance for DCAM and expert routing in DMoE. The self-attention operation is defined as follows:

$$Q = S_{embed}W_Q, K = S_{embed}W_K, V = S_{embed}W_V, \quad (6)$$

$$S_{prior} = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (7)$$

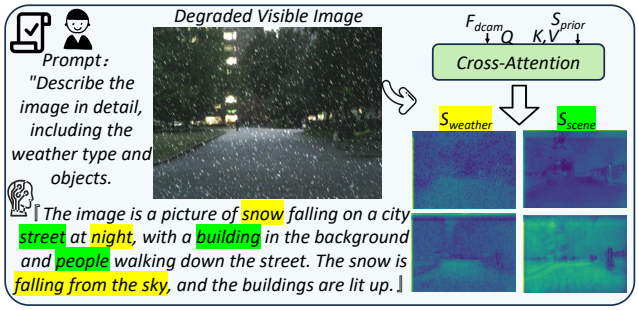

Figure 4: The process of semantic prior extraction and its deep interaction with image features.

where $W_Q, W_K, W_V \in \mathbb{R}^{C\times d_k}$ are learnable projection matrices for the query, key, value, and d_k is the scaling factor.

An example of **the semantic prior extraction is illustrated in Fig. 4**, where the prior consists of two parts: $S_{weather}$, representing weather-aware degradation knowledge, and S_{scene}, capturing the scene features. In subsequent expert routing, image features modulated by DCAM interact with this semantic prior—$S_{weather}$ enhances degradation textures in visible images, while S_{scene} strengthens object information in both infrared and visible features. This demonstrates the semantic prior's deep guidance on image features, highlighting critical regions to enable adaptive expert routing.

Degradation-Aware Channel Attention Module (DCAM) We propose a degradation-aware module that leverages the semantic prior for channel modulation. Specifically, we utilize the semantic prior from DSPE, which encodes both weather-related $S_{weather}$ and scene-related S_{scene} information to identify similar and distinguish different degradation scenarios.

To decompose the semantic prior into degradation prototypes, we first pass S_{prior} through a dimensionality reduction MLP layer $\Phi_m^{II}(\cdot)$ to produce a score vector $s_K \in \mathbb{R}^K$, where each entry corresponds to the activation score of one of the K degradation prototypes. It is formulated as:

$$s_K = \sigma(\mathcal{N}_{layer}(\Phi_m^{II}(\mathcal{P}_{avg}(S_{prior})))), \quad (8)$$

where $\mathcal{P}_{avg}(\cdot)$ denotes average pooling, and $\sigma(\cdot)$ is the Sigmoid activation function.

The score vector s_K is then used to represent the degradation scenario, expressed as a linear combination of the degradation prototypes:

$$D_s = \sum_{i=1}^{K} s_{K_i} \cdot \mathcal{P}_i, \quad (9)$$

where D_s denotes the specific degradation scenario, and $\mathcal{P}_i$ denotes the i-th degradation prototype.

Then, we model the relationship between degradation prototypes and channel responses by representing each degradation prototype as a vector $k_i \in \mathbb{R}^C$, where C is the number of channels in F_{in}. Each vector encodes the response strength of the i-th prototype across the C channels.

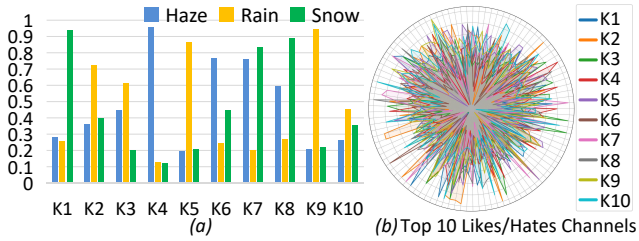

Figure 5: (a) Decomposition of semantic prior into degradation prototypes across different weather types. (b) Radar chart of top 10 activated (outward) and suppressed (inward) channels per degradation prototype.

Consequently, the channel-wise weights can be obtained as:

$$w_c = \sigma\left(\sum_{i=1}^{K} s_{K_i} \cdot k_i\right), \tag{10}$$

where $w_c \in \mathbb{R}^C$ represents the channel-wise weights.

To represent the degradation prototypes, we concatenate the K degradation prototype vectors $k_i \in \mathbb{R}^C$ into a degradation prototype matrix $W_{proto} \in \mathbb{R}^{K\times C}$. This matrix serves as the practical implementation of the degradation prototypes.

To ensure distinct channel response patterns, prototype vectors are initialized orthogonally and normalized to the range $[-1, 1]$, representing activation strengths across channels. These vectors form the prototype matrix W_{proto}, where each row corresponds to a distinct prototype vector. The normalized orthogonal matrix W_{proto} is then treated as a learnable parameter.

In summary, the semantic prior decomposition generates degradation prototypes that encode distinct channel response patterns, enabling degradation-aware channel modulation, formally expressed as:

$$F_{dcam} = \mathcal{N}_{layer}(F_{in}) \odot \sigma\left(s_K W_{proto}\right) + F_{in}, \tag{11}$$

where $\odot$ denotes channel-wise element-wise multiplication.

As shown in Fig. 5 (a), under different weather conditions, the semantic prior is decomposed into a set of degradation prototypes, whose corresponding proportions display both clear distinctions and latent correlations across different scenarios. This demonstrates that the semantic prior effectively captures degradation-aware semantics — a high-level representation of degradtion senarios — which can be further factorized into a weighted composition of multiple degradation prototypes. **As illustrated in Fig. 5 (b)**, our training strategy enables each prototype to learn diverse and distinct channel preference patterns. Such diversity improves the prototype mixture's expressiveness and adaptability to various degradation scenarios.

Degradation-Aware Mixture-of-Experts (DMoE) To implement a degradation-aware MoE network, we establish cross-modal interactions between the semantic prior S_{prior} and the DCAM-weighted feature maps F_{dcam}, generating expert activation scores.

We use a cross-attention mechanism to interact the semantic prior S_{prior} with image features, emphasizing those features related to the degradation scenario. It is expressed as:

$$Q = F_{dcam}W_Q, K = S_{prior}W_K, V = S_{prior}W_V, \tag{12}$$

$$\hat{a} = \mathcal{FL}(Softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V + F_{dcam}), \tag{13}$$

where $\mathcal{FL}(\cdot)$ denotes the flattening operation. The cross-attention output is then processed through a dual-branch layer to reduce dimensionality:

$$F_b^I = \phi\left(\mathcal{N}_{layer}(\mathcal{T}(\hat{a}))\right), F_b^{II} = \mathcal{P}_{avg}\left(\phi(\mathcal{N}_{layer}(\hat{a}))\right), \tag{14}$$

where $\phi(\cdot)$ denotes the GELU activation function, $\mathcal{T}(\cdot)$ represents a linear transformation layer used for dimensionality reduction, and F_b^I, F_b^{II} are then passed through a convolutional layer $\mathcal{C}(\cdot)$ for channel reduction. This is followed by a softmax operation to compute the expert activation scores $w \in \mathbb{R}^{1\times N}$, where N denotes the number of experts:

$$w = \text{Softmax}\left(\mathcal{C}\left(\text{Cat}\left(F_b^I, F_b^{II}\right)\right)\right). \tag{15}$$

Then, F_{dcam} is processed in parallel by a set of $N = 5$ experts, each sharing an identical lightweight architecture. Specifically, each expert comprises a 3×3 convolutional layer $\mathcal{C}_{3\times3}(\cdot)$ followed by a 1×1 convolution $\mathcal{C}_{1\times1}(\cdot)$. The output of this expert ensemble is formulated as:

$$E_i = \mathcal{C}_{1\times1}\left(\mathcal{C}_{3\times3}(F_{dcam})\right), \tag{16}$$

$$F_{dmoe} = \phi\left(\mathcal{N}_{batch}\left(\sum_{i=1}^{N} w_i E_i\right)\right), \tag{17}$$

where E_i denotes the output of the i-th expert, w_i indicates its corresponding activation weight, and $\mathcal{N}_{batch}(\cdot)$ represents batch normalization.

Decoder The enhanced feature F_{dmoe} is then fed into a CNN-based decoder $\mathcal{D}(\cdot)$ for image reconstruction, as formulated below:

$$I_f = \mathcal{D}(F_{dmoe}). \tag{18}$$

Loss Functions

We design loss functions based on the framework from prior work (Zhang et al. 2024). To address the degradation, we adapt it to the fusion-based degradation removal task. Specifically, we use the original undegraded visible image $\tilde{I}_{vi}$ and I_{ir} as ground-truth.The overall image fusion loss L_{fusion} is the sum of the integration L_{inte} and color consistency losses L_{color}:

$$L_{fusion} = L_{inte} + L_{color}. \tag{19}$$

Integration Loss To preserve texture and edge details, we define the integration loss L_{inte}, which ensures the fused image matches the pixel-wise intensity and gradient maxima of the source images. It is formulated as:

$$L_{inte} = \left\|\nabla I_f - \max(\nabla\tilde{I}_{vi}, \nabla I_{ir})\right\|_1 + \left\|I_f - \max(\tilde{I}_{vi}, I_{ir})\right\|_1, \tag{20}$$

where ∇I represents the image gradient, and $\max(\cdot)$ denotes the pixel-wise maximum operation.

Method	Haze				Method	Rain				Method	Snow			
	PSNR↑	SSIM↑	NABF↓	MI↑		PSNR↑	SSIM↑	NABF↓	MI↑		PSNR↑	SSIM↑	NABF↓	MI↑
+ SegMiF					**+ SegMiF**					**+ SegMiF**				
DCMPNet	16.769	1.019	0.0287	1.673	DRSformer	17.308	0.859	0.0719	1.722	HDCWNet	16.513	0.773	0.1149	1.594
CasDyF-Net	15.700	0.907	0.0203	1.779	IDT	16.515	0.608	0.1312	1.305	InvDSNet	16.051	0.612	0.1200	1.575
DehazeFormer	17.051	1.046	0.0222	2.183	NeRD-Rain	16.461	0.585	0.1355	1.289	SnowFormer	16.007	0.616	0.1224	1.606
+ SwinFuse					**+ SwinFuse**					**+ SwinFuse**				
DCMPNet	15.631	0.442	0.0277	1.423	DRSformer	17.000	0.663	0.0225	1.366	HDCWNet	15.518	0.518	0.0710	1.430
CasDyF-Net	16.592	0.743	0.0230	1.911	IDT	17.271	0.826	0.0369	1.246	InvDSNet	15.639	0.533	0.0755	1.403
DehazeFormer	15.558	0.364	0.0237	1.416	NeRD-Rain	17.261	0.830	0.0380	1.238	SnowFormer	15.445	0.499	0.0829	1.440
+ TarDAL					**+ TarDAL**					**+ TarDAL**				
DCMPNet	15.719	0.144	0.0226	0.612	DRSformer	16.146	0.223	0.0310	0.758	HDCWNet	15.839	0.236	0.0578	0.799
CasDyF-Net	15.871	0.181	0.0267	0.705	IDT	16.255	0.279	0.0454	0.728	InvDSNet	15.884	0.254	0.0634	0.764
DehazeFormer	15.716	0.143	0.0212	0.614	NeRD-Rain	16.260	0.283	0.0463	0.733	SnowFormer	15.853	0.254	0.0648	0.794
+ MetaFusion					**+ MetaFusion**					**+ MetaFusion**				
DCMPNet	16.717	1.104	0.0590	1.394	DRSformer	17.058	0.917	0.1201	1.235	HDCWNet	16.210	0.831	0.1784	1.281
CasDyF-Net	15.530	1.040	0.0521	1.646	IDT	15.290	0.561	0.1928	0.903	InvDSNet	15.590	0.657	0.1924	1.183
DehazeFormer	16.940	1.149	0.0472	1.611	NeRD-Rain	15.226	0.535	0.1973	0.886	SnowFormer	15.668	0.669	0.1921	1.214
+ SAGE					**+ SAGE**					**+ SAGE**				
DCMPNet	16.589	1.171	0.0222	1.713	DRSformer	17.964	0.993	0.0704	1.880	HDCWNet	17.267	0.897	0.1176	1.747
CasDyF-Net	13.957	1.025	0.0171	1.969	IDT	16.264	0.665	0.1373	1.490	InvDSNet	16.493	0.712	0.1243	1.793
DehazeFormer	17.260	1.231	0.0157	2.005	NeRD-Rain	16.220	0.641	0.1418	1.481	SnowFormer	16.548	0.716	0.1261	1.752
Ours	**18.325**	**1.302**	**0.0111**	**2.273**	**Ours**	**18.079**	**1.260**	**0.0131**	**2.406**	**Ours**	**17.528**	**1.245**	**0.0151**	**2.491**

Table 1: Quantitative comparison under Strategy I on Hazy, rainy and snowy MSRS dataset.

Color Consistency Loss To align the chrominance with the visible image, we introduce a color consistency loss L_{color}, which is defined as:

$$L_{color} = \left\| C_b^f - C_b^{\tilde{vi}} \right\|_1 + \left\| C_r^f - C_r^{\tilde{vi}} \right\|_1, \qquad (21)$$

where C_b and C_r are the chrominance components in the YCbCr color space. Superscripts f and $\tilde{vi}$ refer to the fused and clean visible images, respectively.

Experiments

Experimental Configurations

Datasets The datasets used in our experiments are derived from two widely adopted IVF benchmarks: MSRS (Tang et al. 2022) and FMB (Liu et al. 2023a), both providing clean and well-aligned visible-infrared image pairs. MSRS contains 1,524 pairs (480×640), split into 1,163 for training and 361 for testing. FMB includes 1,500 aligned pairs (600×800), with 1,220 for training and 280 for testing.

To simulate adverse weather conditions, we generated three degraded versions of the full MSRS and FMB datasets by applying haze, rain, and snow effects separately to all visible images. Specifically, haze was synthesized using the Depth-Anything model (Yang et al. 2024c) and the atmospheric scattering model (Zhang, Ding, and Sharma 2017); rain was produced by combining random noise and motion blur; and snow was generated using the imgaug library (Jung et al. 2020). This process resulted in three complete degraded variants of the original datasets, leading to a total of 9,072 registered image pairs (7,149 for training and 1,923 for testing), evenly distributed among the three degradation types (1:1:1).

Implementation Details We set the initial learning rate to 6×10^{-5}, with a batch size of 15, and train the model for 1,000 epochs in total. The learning rate is scheduled using a combination of warm-up and polynomial decay, and we employ the Adam optimizer for optimization. All experiments are conducted on NVIDIA GeForce RTX 4090 GPU with 24GB memory and the AMD Ryzen 9 7950X 16-Core Processor CPU.

Quantitative Comparison

To validate the superiority of our degradation-aware fusion framework across multiple degradation scenarios, we compare it with two strategies: **Strategy I**, which applies dedicated restoration models for each degradation type (haze, rain, snow) followed by baseline fusion; and **Strategy II**, which uses all-in-one restoration models to preprocess degraded inputs before baseline fusion. *In contrast, our method directly fuses original degraded inputs without separate restoration, yielding more effective and streamlined performance.*

We conduct evaluations on 1,923 synthetically degraded image pairs from the MSRS and FMB datasets, benchmarking against five SOTA fusion methods (SAGE (Wu et al. 2025), SegMiF (Liu et al. 2023a), MetaFusion (Zhao et al. 2023a), TarDAL (Liu et al. 2022), SwinFuse (Wang et al. 2022)) combined with restoration networks. These include dedicated restoration models for dehazing (DCMPNet (Zhang, Zhou, and Li 2024), CasDyF-Net (Wang and He 2024), DehazeFormer (Song et al. 2023)), deraining (DRSformer (Chen et al. 2023b), IDT (Xiao et al. 2022), NeRD-Rain (Chen, Pan, and Dong 2024)), desnowing (HDCWNet (Chen et al. 2021), InvDSNet (Quan et al. 2023), SnowFormer (Chen et al. 2022)), and all-in-one restoration

Figure 6: Qualitative comparison on degraded MSRS and FMB datasets. Rows 1-3: MSRS dataset, organized by weather conditions. Rows 4-6: FMB dataset, also grouped by weather.

models (DTMR (Patil et al. 2023), MPMF-Net (Wen et al. 2025), WGWS-Net (Zhu et al. 2023)), all paired with fixed fusion modules. Performance is evaluated by PSNR, SSIM, Nabf, and MI metrics.

Quantitative results under Strategy I on three degradation types of the MSRS dataset are presented in Table 1, where ↑ and ↓ denote preferable higher or lower values, and **bold** and underline highlight the best and second-best results, respectively. Our method achieves the best performance in most cases, as evidenced by the quantitative metrics. The complete quantitative results comparing both Strategy I and II on the degraded MSRS and FMB datasets are provided in the extended version.

Qualitative Comparison

To identify the optimal restoration model for each baseline fusion method, we select the restoration model achieving the highest PSNR for each degradation type and dataset, based on the complete quantitative comparison results (see extended version), denoted as BestR. Using BestR, we construct fusion pipelines for baseline methods and qualitatively compare them with our approach, as illustrated in Fig. 6. Our method consistently delivers clearer, higher-quality fused images across diverse weather conditions and datasets, outperforming SOTA techniques.

Ablation Study

Ablation studies in Table 2 validate the effectiveness of DCAM and DMoE, with optimal results only when both are used. Removing DCAM disables channel-wise modulation, while removing DMoE makes routing purely image-driven, leading to performance drops (first row) and expert collapse—where only one expert remains active (visualizations provided in extended version). These results underscore the necessity of semantic prior guidance for degradation-aware routing and overall performance.

DCAM	DMoE	PSNR↑	SSIM↑	NABF↓	MI↑
×	×	16.265	1.144	0.0142	2.006
×	✓	15.350	1.083	0.0148	2.159
✓	×	17.009	1.194	0.0135	2.101
✓	✓	**17.977**	**1.269**	**0.0131**	**2.390**

Table 2: Average ablation results across haze, rain, and snow on the MSRS dataset.

Conclusion

In this paper, we propose MdaIF, a robust one-stop multi-degradation-aware image fusion framework guided by the semantic prior from a pre-trained VLM. Unlike general IVF methods that overlook visible image degradation, MdaIF adapts fusion via two core modules: DCAM, which uses degradation prototype decomposition to modulate channel-wise features, and DMoE, enabling adaptive expert routing under adverse conditions. Compared to fixed-architecture IVF and prompt-guided fusion methods—which rely on either rigid networks or ground-truth degradation labels—our framework eliminates such dependencies by leveraging the semantic prior. MdaIF outperforms SOTA methods under varied degradations, with strong efficacy and generalization.

Acknowledgments

This research is partially supported by the National Natural Science Foundation of China under Grant 62206321; and the Science and Technology Program of Hunan Province under Grant 2024RC3108.

References

Bao, F.; Wang, X.; Sureshbabu, S. H.; Sreekumar, G.; Yang, L.; Aggarwal, V.; Boddeti, V. N.; and Jacob, Z. 2023. Heat-assisted detection and ranging. *Nature*, 619(7971): 743–748.

Cao, Z.; Zhong, Y.; Wang, Z.; and Deng, L.-J. 2025. MMAIF: Multi-task and Multi-degradation All-in-One for Image Fusion with Language Guidance. *arXiv preprint arXiv:2503.14944*.

Chen, J.; Ding, J.; Yu, Y.; and Gong, W. 2023a. THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor. *Neurocomputing*, 527: 71–82.

Chen, S.; Ye, T.; Liu, Y.; Chen, E.; Shi, J.; and Zhou, J. 2022. SnowFormer: Scale-aware Transformer via Context Interaction for Single Image Desnowing. *arXiv preprint arXiv:2208.09703*.

Chen, W.-T.; Fang, H.-Y.; Hsieh, C.-L.; Tsai, C.-C.; Chen, I.; Ding, J.-J.; Kuo, S.-Y.; et al. 2021. ALL Snow Removed: Single Image Desnowing Algorithm Using Hierarchical Dual-Tree Complex Wavelet Representation and Contradict Channel Loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4196–4205.

Chen, X.; Li, H.; Li, M.; and Pan, J. 2023b. Learning a Sparse Transformer Network for Effective Image Deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5896–5905.

Chen, X.; Pan, J.; and Dong, J. 2024. Bidirectional Multi-Scale Implicit Neural Representations for Image Deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

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

Fu, X.; Huang, J.; Ding, X.; Liao, Y.; and Paisley, J. 2017. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6): 2944–2956.

Jiang, C.; Ren, H.; Yang, H.; Huo, H.; Zhu, P.; Yao, Z.; Li, J.; Sun, M.; and Yang, S. 2024. M2FNet: Multi-modal fusion network for object detection from visible and thermal infrared images. *International Journal of Applied Earth Observation and Geoinformation*, 130: 103918.

Judd, K. M.; Thornton, M. P.; and Richards, A. A. 2019. Automotive sensing: Assessing the impact of fog on LWIR, MWIR, SWIR, visible, and lidar performance. In *Infrared Technology and Applications XLV*, volume 11002, 322–334. SPIE.

Jung, A. B.; Wada, K.; Crall, J.; Tanaka, S.; Graving, J.; Reinders, C.; Yadav, S.; Banerjee, J.; Vecsei, G.; Kraft, A.; Rui, Z.; Borovec, J.; Vallentin, C.; Zhydenko, S.; Pfeiffer, K.; Cook, B.; Fernández, I.; De Rainville, F.-M.; Weng, C.-H.; Ayala-Acevedo, A.; Meudec, R.; Laporte, M.; et al. 2020. imgaug. https://github.com/aleju/imgaug. Online; accessed 01-Feb-2020.

Li, H.; Yang, Z.; Zhang, Y.; Jia, W.; Yu, Z.; and Liu, Y. 2025. MulFS-CAP: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Q.; Lu, L.; Li, Z.; Wu, W.; Liu, Z.; Jeon, G.; and Yang, X. 2019. Coupled GAN with relativistic discriminators for infrared and visible images fusion. *IEEE Sensors Journal*, 21(6): 7458–7467.

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.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; and Fan, X. 2023a. Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for Image Fusion and Segmentation. In *International Conference on Computer Vision*.

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

Liu, Z.; Liu, J.; Zhang, B.; Ma, L.; Fan, X.; and Liu, R. 2023b. PAIF: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *Proceedings of the 31st ACM international conference on multimedia*, 3706–3714.

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

Patil, P. W.; Gupta, S.; Rana, S.; Venkatesh, S.; and Murala, S. 2023. Multi-weather Image Restoration via Domain Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21696–21705.

Quan, Y.; Tan, X.; Huang, Y.; Xu, Y.; and Ji, H. 2023. Image desnowing via deep invertible separation. *IEEE Transactions on Circuits and Systems for Video Technology*.

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 *International conference on machine learning*, 8748–8763. PmLR.

Ren, X.; Meng, F.; Hu, T.; Liu, Z.; and Wang, C. 2018. Infrared-visible image fusion based on convolutional neural

networks (CNN). In *International Conference on Intelligent Science and Big Data Engineering*, 301–307. Springer.

Song, Y.; He, Z.; Qian, H.; and Du, X. 2023. Vision Transformers for Single Image Dehazing. *IEEE Transactions on Image Processing*, 32: 1927–1941.

Tang, L.; Yuan, J.; Zhang, H.; Jiang, X.; and Ma, J. 2022. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*.

Velázquez, J. M. R.; Khoudour, L.; Saint Pierre, G.; Duthon, P.; Liandrat, S.; Bernardin, F.; Fiss, S.; Ivanov, I.; and Peleg, R. 2022. Analysis of thermal imaging performance under extreme foggy conditions: Applications to autonomous driving. *Journal of imaging*, 8(11): 306.

Wang, H.; Zhang, H.; Yi, X.; Xiang, X.; Fang, L.; and Ma, J. 2024. Terf: Text-driven and region-aware flexible visible and infrared image fusion. In *Proceedings of the 32nd ACM international conference on multimedia*, 935–944.

Wang, Y.; and He, B. 2024. CasDyF-Net: Image Dehazing via Cascaded Dynamic Filters. In *2024 IEEE 8th International Conference on Vision, Image and Signal Processing (ICVISP)*, 1–8. IEEE.

Wang, Z.; Chen, Y.; Shao, W.; Li, H.; and Zhang, L. 2022. SwinFuse: A Residual Swin Transformer Fusion Network for Infrared and Visible Images. *IEEE Transactions on Instrumentation and Measurement*, 1–1.

Wen, Y.; Gao, T.; Zhang, J.; Li, Z.; and Chen, T. 2025. Multi-axis prompt and multi-dimension fusion network for all-in-one weather-degraded image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8323–8331.

Wu, G.; Liu, H.; Fu, H.; Peng, Y.; Liu, J.; Fan, X.; and Liu, R. 2025. Every SAM Drop Counts: Embracing Semantic Priors for Multi-Modality Image Fusion and Beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xiao, J.; Fu, X.; Liu, A.; Wu, F.; and Zha, Z.-J. 2022. Image De-raining Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

Yang, B.; Hu, Y.; Liu, X.; and Li, J. 2024a. CEFusion: An Infrared and Visible Image Fusion Network Based on Cross-Modal Multi-Granularity Information Interaction and Edge Guidance. *IEEE Transactions on Intelligent Transportation Systems*, 25(11): 17794–17809.

Yang, H.; Pan, L.; Yang, Y.; and Liang, W. 2024b. Language-driven All-in-one Adverse Weather Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24902–24912.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024c. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *CVPR*.

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 (CVPR)*.

Zhang, H.; Cao, L.; Zuo, X.; Shao, Z.; and Ma, J. 2025. OmniFuse: Composite Degradation-Robust Image Fusion with Language-Driven Semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H.; Xu, H.; Tian, X.; Jiang, J.; and Ma, J. 2021. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76: 323–336.

Zhang, H.; Zuo, X.; Jiang, J.; Guo, C.; and Ma, J. 2024. 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(8): 10535–10554.

Zhang, Y.; Ding, L.; and Sharma, G. 2017. Hazerd: an outdoor scene dataset and benchmark for single image dehazing. In *2017 IEEE international conference on image processing (ICIP)*, 3205–3209. IEEE.

Zhang, Y.; Song, B.; Du, X.; and Guizani, M. 2018. Vehicle Tracking Using Surveillance With Multimodal Data Fusion. *IEEE Transactions on Intelligent Transportation Systems*, 19(7): 2353–2361.

Zhang, Y.; Zhou, S.; and Li, H. 2024. Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2846–2855.

Zhao, W.; Xie, S.; Zhao, F.; He, Y.; and Lu, H. 2023a. MetaFusion: Infrared and Visible Image Fusion via Meta-Feature Embedding from Object Detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13955–13965.

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

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.

Zhu, Y.; Wang, T.; Fu, X.; Yang, X.; Guo, X.; Dai, J.; Qiao, Y.; and Hu, X. 2023. Learning Weather-General and Weather-Specific Features for Image Restoration Under Multiple Adverse Weather Conditions. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.