HDCFN: Haze Distribution-aware Cross-modal Fusion Network for Infrared-guided Dense Haze Removal in UAVs

Junwei Zhao
China Telecom Cloud Computing Research Institute, China
zhaojw1@chinatelecom.cn

Qianchun Luo
China Telecom Unmanned Technology Jiangsu Co., Ltd., China
sqlqc@189.cn

Shiliang Zhang
School of Computer Science, Peking University, China
slzhang.jdl@pku.edu.cn

Shen Gao
China Telecom Cloud Computing Research Institute, China
gaos13@chinatelecom.cn

Jie Wu
China Telecom Cloud Computing Research Institute, China
wujie@chinatelecom.cn

Abstract

In UAV applications, dense haze severely obscures small ground-level objects, hindering the recovery of fine details. Existing visible-only dehazing methods struggle with such dense occlusions, while infrared imaging lacks color and fine texture information. To address these limitations, we propose the Haze Distribution-aware Cross-modal Fusion Network (HDCFN). HDCFN features two key components: (i) an infrared-guided multiscale feature enhancement framework that integrates haze-resistant structural cues from infrared modality with visible features across coarse to fine, improving the recovery of small objects, and (ii) a haze distribution-aware cross-modal fusion module that adaptively prioritizes relevant information from each modality according to haze density. This framework effectively combines the complementary strengths of visible and infrared imaging for dense haze removal. Extensive experiments on multiple public datasets show that HDCFN outperforms state-of-the-art dehazing and fusion methods, yielding higher-quality and more detailed images.

CCS Concepts

• **Artificial intelligence** → **Computer vision representations; Vision for robotics**.

Keywords

Multimodal Representation Learning and Perception, Edge Intelligence, UAV

ACM Reference Format:

Junwei Zhao, Qianchun Luo, Shiliang Zhang, Shen Gao, and Jie Wu. 2025. HDCFN: Haze Distribution-aware Cross-modal Fusion Network for Infrared-guided Dense Haze Removal in UAVs. In *Proceedings of the 33st ACM International Conference on Multimedia (MM '25), Oct 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/11.1145/3381783.3612099

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/11.1145/3381783.3612099

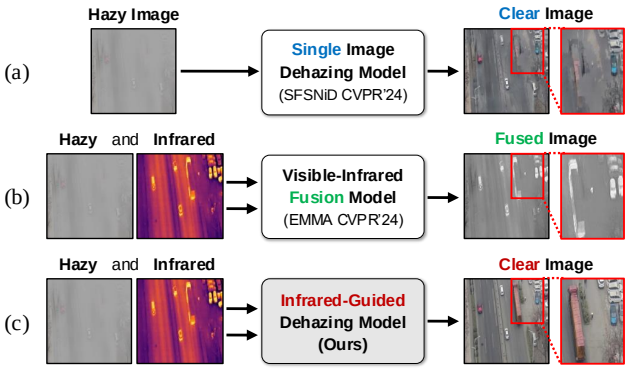

Figure 1: Comparison of existing methods and the proposed method. (a) Single-image dehazings [3] using visible-only model struggles with *dense haze* due to the lack of structural guidance. (b) Visible-infrared fusion [48] integrates both modalities but focuses on general-purpose fusion, resulting in inadequate color and details restoration for haze removal. (c) The proposed infrared-guided dehazing method leverages the strengths of both modalities, achieving superior restoration with clearer, more accurate outputs.

1 Introduction

In high-humidity weather, visual imagery is degraded by haze, resulting in reduced visibility and clarity [12]. This issue is particularly pronounced in UAV scenarios, such as aerial surveillance and remote sensing, where precise visual representation is critical for decision-making [34, 41]. In these scenarios, ground targets are often small and distant, while the surrounding environment is cluttered [14, 38]. Dense haze can completely obscure these targets, complicating tasks such as feature reconstruction and detection.

Visible light imaging is less effective in such conditions, as dense haze significantly reduces contrast and increases scattering [11, 20]. Recent deep learning-based dehazing methods have shown promise in restoring haze-free images from degraded visible light inputs [8]. Most of these methods rely on prior assumptions or generative models [9, 49]. However, in dense haze, these methods struggle to recover details due to the lack of reliable visual cues, resulting in incomplete restoration and visible artifacts as shown in Fig. 1(a) [29, 40]. These limitations are particularly challenging in

drone-captured aerial imagery, where small, fully obscured objects are difficult to infer from surrounding cues.

Thermal infrared imaging offers advantages in haze penetration and the preservation of structural cues, but it lacks fine texture and color fidelity [1]. These limitations highlight the need for multimodal fusion, which combines visible and infrared data to leverage the complementary strengths of both modalities [44]. However, existing visible-infrared fusion methods (Fig. 1(b)) typically optimize for general-purpose fusion objectives, rather than addressing the specific challenges related to haze removal [22, 36]. This misalignment between general fusion objectives and the specific demands of dehazing limits their effectiveness in real-world UAV operations, where the dynamic nature and spatial variability of haze necessitate adaptive, context-aware fusion techniques.

To address the above challenges of small-scale object recovery and the removal of dynamic haze with varying spatial distribution, we propose a haze distribution-aware cross-modal fusion network. Specifically, we design a multiscale feature enhancement framework that leverages infrared structural features to guide the refinement of visible features from coarse to fine. This hierarchical framework improves small object recovery and enhances fine details. Furthermore, we propose a cross-modal fusion module that adaptively integrates content from different modalities based on the density of haze spatial distribution. In regions with high haze density, the model prioritizes infrared features to better preserve structural information, effectively mitigating haze-induced degradation in visible images. This adaptive fusion improves the network's robustness to varying haze conditions, ensuring clearer and more accurate outputs for UAV applications as shown in Fig. 1(c).

We conducted extensive experiments on publicly available UAV-captured datasets, demonstrating that the integration of infrared maps significantly enhances visible image dehazing. Our method achieves 8.1% and 7.9% PSNR improvements on the VTUAV and CART datasets, respectively, compared to state-of-the-art methods. Notably, integrating infrared maps yields more pronounced improvements in dense haze conditions. Furthermore, validation on our real-world UAV-captured data further demonstrates the effectiveness and robustness of the proposed method.

Our contributions can be summarized as follows:

- To the best of our knowledge, it is the first work to utilize infrared features to guide visible light dehazing in UAV scenarios, where small objects are prone to dense occlusions and dynamic haze.
- We propose the Haze Distribution-aware Cross-modal Fusion Network (HDCFN), incorporating an infrared-guided multi-scale feature enhancement framework and a haze distribution-aware cross-modal adaptive fusion module.
- Extensive experiments on multiple public datasets demonstrate substantial performance improvements over SoTA dehazing methods. Furthermore, we validate the method using real-world data captured by our drones, highlighting its practical applicability.

2 Related Works

2.1 Image Dehazing

Image dehazing aims to recover clear images from haze-degraded inputs. Early methods relied on physical priors such as dark channel prior [10] and atmospheric scattering models [25]. While effective in certain scenarios, these approaches often fail under complex haze conditions due to their reliance on fixed assumptions [7]. Recent advances in deep learning have significantly improved dehazing performance by learning representations directly from data [2, 23, 39]. Notable models, such as DehazeFlow [15], IR-SDE [24], DehazeFormer [30], and DCMPNet [45], demonstrate enhanced performance through advanced network architectures and learning strategies. However, single image dehazing typically relies on generative models, which struggle in extremely hazy conditions, as haze severely obscures key scene details.

The integration of infrared modality to assist in dehazing visible light images remains in its early stages [6, 16]. Yu et al. [42] propose an encoder-decoder framework utilizing infrared data to enhance hazy RGB images, but it faces limitations in dense haze and UAV scenarios. This highlights the need for more robust and adaptive methods for extreme haze conditions.

2.2 Visible-Infrared Fusion

Visible-infrared fusion methods leverage the complementary strengths of visible and infrared images, with deep learning-based approaches becoming prominent for refining cross-modality features [31, 33, 37]. Recent advancements in deep learning-based fusion methods [17, 32, 44] primarily focus on feature-level integration, aiming to combine the strengths of both modalities. These approaches generally emphasize feature extraction or contrast enhancement, with limited exploration in dehazing [18]. In contrast to existing visible-infrared fusion methods, this work introduces infrared guidance for haze removal, particularly in UAV scenarios.

3 Preliminary

The visible light imaging process can be modeled using the atmospheric scattering model [27], as illustrated in Fig. 2(a), and is formulated as:

$$I_{\mathrm{RGB}}(x) = J(x) \cdot t_{\mathrm{RGB}}(x) + A \cdot (1 - t_{\mathrm{RGB}}(x)), \tag{1}$$

where $I_{\mathrm{RGB}}(x)$ represents the observed hazy image, $J(x)$ is the clear image to be recovered, $t_{\mathrm{RGB}}(x) = e^{-\beta_{\mathrm{RGB}} d(x)}$ denotes the transmission map indicating the fraction of light reaching the camera, and A is the global atmospheric light. Here, β_{RGB} is the atmospheric scattering coefficient and $d(x)$ is the scene depth.

The thermal infrared imaging process, as depicted in Fig. 2(b) and described in [21], can be expressed as:

$$I_{\mathrm{IR}}(x) = I_{\mathrm{s}}(x) \cdot e^{-\beta_{\mathrm{IR}} d(x)} + A \cdot (1 - e^{-\beta_{\mathrm{IR}} d(x)}), \tag{2}$$

where $I_{\mathrm{IR}}(x)$ denotes the observed infrared intensity, $I_{\mathrm{s}}(x)$ represents the intrinsic infrared reflectance of the scene, β_{IR} is the infrared scattering coefficient.

The difference in scattering between visible and infrared light arises from Rayleigh scattering [28], which occurs when light interacts with particles smaller than its wavelength. As summarized in Tab. 1, shorter wavelengths (e.g., visible light, 400~780 nm) experience significantly stronger scattering than longer wavelengths (e.g., thermal infrared, 8~14 μm). Consequently, the relationship between the scattering coefficients is:

$$\beta_{\mathrm{IR}} \ll \beta_{\mathrm{RGB}}. \tag{3}$$

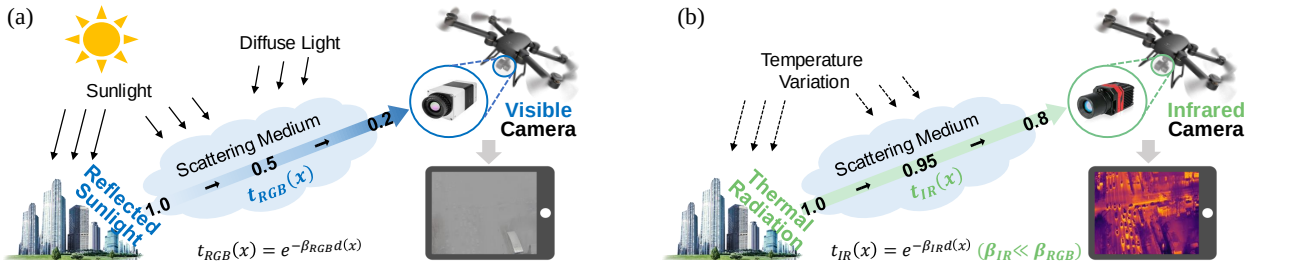

Figure 2: Comparison of visible and thermal infrared imaging in dense haze condition. (a) Atmospheric scattering model illustrates the severe attenuation of visible light due to scattering and absorption. (b) Thermal infrared imaging model highlights the reduced impact of scattering on longer wavelengths, providing complementary structural details to assist in dehazing.

Table 1: Transmittance of visible light and thermal infrared across different levels of atmospheric haze.

Modality	Wavelength	Transmittance at various haze levels				
Visible Light	400~780 nm	1.00	0.80	0.60	0.40	0.20
Thermal Infrared	8~14 μm	1.00	0.97	0.96	0.93	0.80

This wavelength-dependent behavior leads to rapid attenuation of visible light in dense hazy conditions. We calculate the transmittance of both visible light and thermal infrared at the same haze levels, with the results recorded in Tab. 1. The corresponding haze effect visualizations are presented in Fig. 3. The comparison shows that infrared imaging maintains a higher transmission rate, enabling clearer capture of structural details.

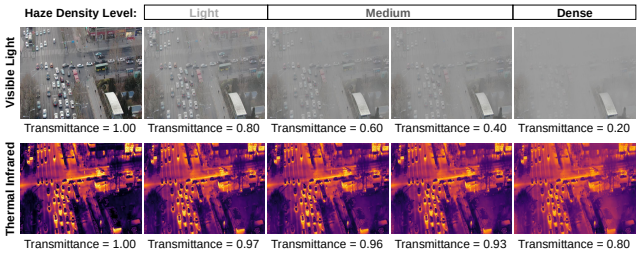

Figure 3: Visualization of visible light and thermal infrared images at different transmittance levels.

4 Methodology

4.1 Formulation

Our objective is to enhance aerial image dehazing by leveraging the complementary properties of visible and infrared images. Visible images contain essential color and texture features for naturalistic outputs, but suffer from severe degradation under dense haze, losing crucial visual cues. Conversely, infrared images, while resistant to haze and capable of capturing object shapes and contours, lack color fidelity and texture details.

To leverage these strengths, we propose an infrared-guided cross-modal network to fuse features from both modalities for robust dehazing. Our goal is to produce a dehazed visible image $\hat{V}$ from a spatially aligned input pair consisting of a visible image V and an infrared image I. We formulate this as the following optimization:

$$\hat{V} = \arg\min_{\theta} \mathcal{L}(\mathcal{F}_\theta(V, I), V_{\text{gt}}), \tag{4}$$

where $\mathcal{F}_\theta$ represents our network parameterized by θ, and V_{gt} is the ground-truth dehazed image. The loss function $\mathcal{L}$ is designed to guide $\hat{V}$ to retain structural features from infrared while preserving visible color and texture.

4.2 Infrared-guided Multiscale Feature Enhancement (IMFE)

As illustrated in Fig. 4, this framework extracts multiscale features from both visible and infrared inputs to capture details at various resolutions, facilitating the recovery of small targets in UAV scenarios. By combining color and texture cues from visible images with the structural robustness of infrared features, it ensures robust feature representation under haze conditions, which is critical for detecting small, distant, or obscured objects.

The initial feature extraction is performed as:

$$F_V^0 = f_V(V), \quad F_I^0 = f_I(I), \tag{5}$$

where f_V and f_I are convolutional blocks, and $F_V^0, F_I^0 \in \mathbb{R}^{C\times H\times W}$ represents the initial visible and infrared features. These features are progressively downsampled:

$$F_V^{l+1} = d_s(F_V^l), \quad F_I^{l+1} = d_s(F_I^l), \quad l = 0, 1, .., L, \tag{6}$$

where $d_s(\cdot)$ denotes a downsampling operation (e.g., max-pooling). At each scale l, the infrared features guide the visible features through a channel attention module. The attention weights are computed as:

$$\alpha_c^l = \sigma(W_3^l \cdot [\text{ReLU}(W_1^l \cdot g_i^l), \text{ReLU}(W_2^l \cdot g_v^l)]), \tag{7}$$

where $g_i^l(g_v^l) \in \mathbb{R}^C$ is the global average pooling of $F_I^l(F_V^l)$, and $W_1^l, W_2^l \in \mathbb{R}^{C\times C}$, $W_3^l \in \mathbb{R}^{C\times 2C}$ are learnable parameters, $\sigma(\cdot)$ is the sigmoid activation, and $[\,,\,]$ denotes the concatenation operation. The enhanced features at each scale are obtained by applying the channel attention to infrared features and integrating them with visible features as:

$$F_E^l = F_V^l + \alpha_c^l \cdot F_I^l, \tag{8}$$

where α_c^l denotes the channel attention weights. In the upsampling stage, transposed convolutions restore spatial resolution, while skip connections integrate high-resolution features from earlier downsampling layers. This hierarchical fusion framework enhances

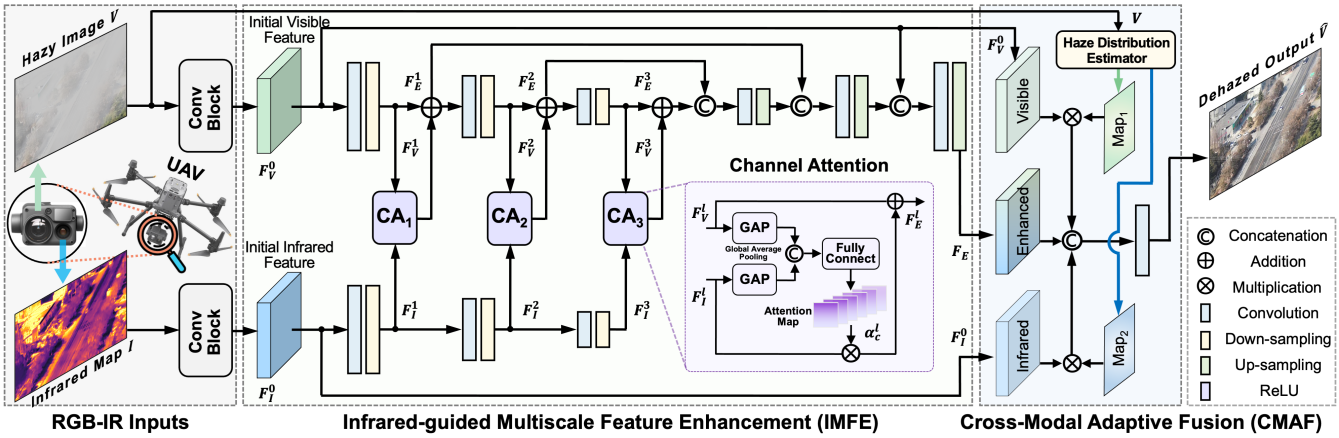

Figure 4: Illustration of the proposed network. The inputs are from an RGB-IR dual-camera UAV. The hierarchical framework enhances the recovery of small objects and fine-grained details by using infrared structural features to guide the improvement of degraded visible features across multiple scales. The haze distribution estimator adapts the weighting of visible and infrared features based on regional haze densities, ensuring effective fusion in dynamic UAV motion scenarios.

small target recovery by effectively complementing visible details with infrared structural features at each scale.

4.3 Haze Distribution-aware Cross-Modal Adaptive Fusion (CMAF)

As illustrated in Fig. 4, the haze distribution-aware Cross-Modal Adaptive Fusion (CMAF) module integrates visible and infrared features by adaptively accounting for spatial haze variations in UAV scenarios. The module leverages a haze density map estimated by the Haze Distribution Estimator (HDE), guiding the feature fusion process across different modalities for context-aware processing.

The haze density map $\mathcal{M}$ is computed by the HDE, which utilizes multilayer deformable convolutions to extract haze-related features from the visible image V, enabling adaptive focus on spatially varying and irregular haze patterns. Given a deformable kernel of K sampling locations, the weight and offset for the k-th location are denoted as ω_k and z_k, respectively. The output feature map $F_D \in \mathbb{R}^{C\times H\times W}$ at each position $z = (h, w)$ can be obtained by:

$$F_D(z) = \sum_{k=1}^{K} \omega_k \cdot V(z + z_k + \Delta z_k) \cdot \Delta m_k, \tag{9}$$

where Δz_k and Δm_k are the learnable offset and modulation scalar for the k-th location, respectively.

Multiscale convolutions are applied to capture features at various spatial resolutions, allowing the model to effectively process both fine-grained details and larger-scale structures in F_D. The feature map is obtained by:

$$F_M = [\mathrm{Conv}_{d_1}(F_D), \mathrm{Conv}_{d_2}(F_D), \mathrm{Conv}_{d_3}(F_D)], \tag{10}$$

where d_n represents convolution kernels with different sizes.

Spatial attention is adopted to emphasize relevant features by combining global average pooling (GAP) and global max pooling (GMP) along the channel axis, allowing the model to focus on the most informative regions:

$$\mathcal{M}(F_M) = \sigma(\mathrm{Conv}([\mathrm{GAP}(F_M), \mathrm{GMP}(F_M)])) . \tag{11}$$

The generated haze density map $\mathcal{M}$ is then used to guide the fusion of visible, infrared, and enhanced features, adapting the fusion weights based on haze intensity:

$$\hat{V} = \mathrm{Conv}([\mathcal{M}\cdot F_I^0,\ (1-\mathcal{M})\cdot F_V^0,\ \eta\cdot F_E]) , \tag{12}$$

where F_I^0, F_V^0, and F_E are the initial infrared, visible, and enhanced feature maps, and η is a scalar factor.

The CMAF module uses the haze density map from the HDE to adaptively guide feature fusion, adjusting fusion weights according to haze distribution. This enables the model to handle spatially varying haze, enhancing feature fusion robustness in UAV imagery.

4.4 Model Training

Our model is trained with a composite loss function $\mathcal{L}$, designed to retain structural fidelity and fine details. The total loss combines a reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ and a cross-modal perception loss $\mathcal{L}_{\mathrm{perc}}$:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}}(\hat{V}, V_{\mathrm{gt}}) + \lambda \cdot \mathcal{L}_{\mathrm{perc}}(F_V^0, F_I^0) , \tag{13}$$

where λ is set to 0.1. The $\mathcal{L}_{\mathrm{rec}}$ is defined as the Mean Absolute Error (MAE) between the predicted dehazed image $\hat{V}$ and the ground-truth image V_{gt}:

$$\mathcal{L}_{\mathrm{rec}} = \|\hat{V} - V_{\mathrm{gt}}\|_1 . \tag{14}$$

The $\mathcal{L}_{\mathrm{perc}}$ is defined as the Mean Squared Error (MSE) between initial visible and infrared feature maps F_V^0 and F_I^0, promoting alignment of structural content from infrared features with visible representations at an early stage:

$$\mathcal{L}_{\mathrm{perc}} = \|F_V^0 - F_I^0\|_2^2 . \tag{15}$$

This combined objective guides the model to generate a dehazed output $\hat{V}$ that retains details from visible input while integrating structural guidance from infrared features, enhancing dehazing robustness and perceptual coherence.

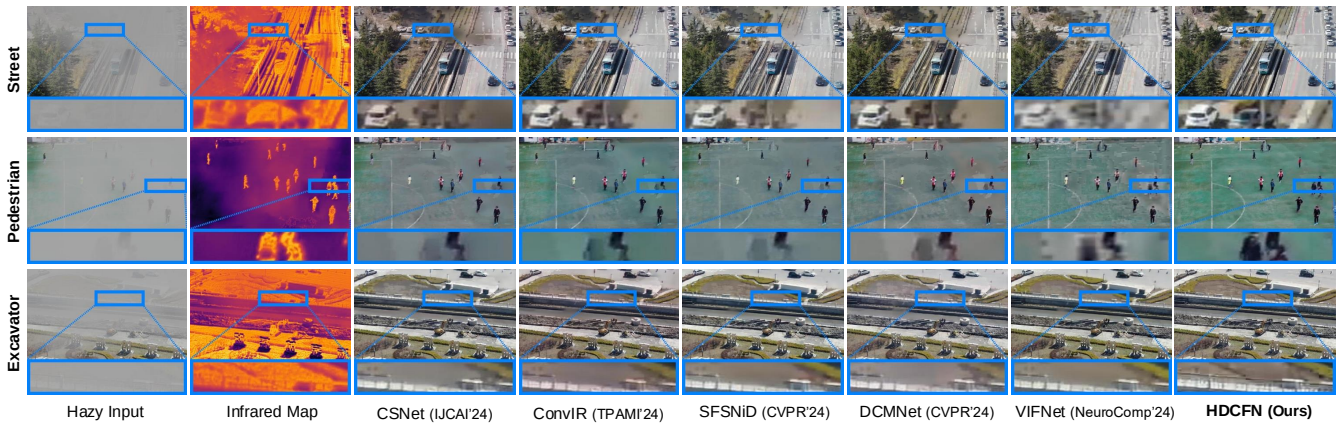

Figure 5: Visualization of partial experimental results on the VTUAV dataset.

Table 2: Comparative results on the VTUAV dataset with SoTA dehazing methods (best results in bold, second-best underlined).

Method	Citation	Modality	Bike		Street		Car		Excavator		Pedestrian		Train		Truck		**Average**	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
CSNet	IJCAI'24	RGB	25.21	0.884	25.48	0.865	25.36	0.831	25.36	0.865	27.43	0.894	25.15	0.889	27.47	0.903	25.92	0.876
ConvIR	TPAMI'24	RGB	28.07	0.891	28.29	0.869	28.16	0.832	29.13	0.872	29.32	0.892	27.45	0.891	30.16	0.907	28.64	0.879
SFSNiD	CVPR'24	RGB	26.41	0.885	25.07	0.859	26.38	0.828	26.68	0.863	27.38	0.893	24.49	0.881	26.05	0.890	26.07	0.871
DCMNet	CVPR'24	RGB+Depth	28.53	0.893	29.13	0.871	28.79	0.835	29.08	0.874	29.14	0.895	29.46	0.889	28.52	0.907	28.95	0.880
VIFNet	NeuroComp'24	RGB+IR	28.79	0.896	29.04	0.871	29.15	0.843	28.67	0.864	28.72	0.895	31.06	0.894	29.03	0.912	29.21	0.881
Ours	ACM MM'25	RGB+IR	**30.56**	**0.905**	**31.69**	**0.913**	**31.25**	**0.914**	**31.42**	**0.921**	**31.51**	**0.916**	**33.32**	**0.925**	**31.24**	**0.933**	**31.57**	**0.918**
Gain (%)	–	–	6.1%	1.0%	8.8%	4.8%	7.2%	8.4%	7.9%	5.4%	7.5%	2.3%	7.3%	3.5%	3.6%	2.3%	**8.1%**	**4.2%**

5 Experiment

5.1 Experimental Settings

Datasets The experiments are conducted on two publicly available real-world synchronized RGB-T datasets and our original UAV-captured dataset. The haze generation algorithm is based on the depth-assisted atmospheric scattering model [20]. A key advantage of using depth information in haze synthesis is that it introduces spatial variation in haze density, which is essential for generating realistic hazy images. Below is an overview of the datasets.

(i) The *VTUAV* [43] dataset provides visible and thermal imagery for UAV-based vision across diverse environments. The dataset encompasses a wide range of scenes featuring objects such as excavators, pedestrians, streets, bikes, cars, trucks, and trains.

(ii) The *CART* [13] dataset offers high-resolution paired RGB-T images captured in natural environments, covering a diverse range of terrains such as rivers, bridges, rocks, and lakes.

(iii) The *CityUAV* dataset is collected using our UAV equipped with a Zenmuse H20T in foggy urban environments, covering diverse city scenes such as streets, buildings, vehicles, parks, bridges, and lakes. The H20T features a visible-light camera with a resolution of 1920×1080 at 30 fps and a thermal infrared camera with a resolution of 640×512 at 30 Hz. The two data streams are time-synchronized, and spatial alignment is achieved using registration software, with both RGB and infrared images resized to 512×512.

Implementation The proposed method is implemented using PyTorch, with experiments conducted on NVIDIA A100 GPUs. The models are optimized using the Adam optimizer (β_1 = 0.9 and β_2 = 0.999). The initial learning rate is set to 2e-4 and gradually reduced to 1e-6 with the cosine annealing. Batch size is configured to 2. For both the baselines and our models, training images (including RGB and infrared maps) are resized to 512×512 as inputs.

5.2 Comparison with SoTA Dehazing Methods

We conducted comparative experiments with five advanced image dehazing methods: CSNet [4], ConvIR [5], SFSNiD [3], DCMNet [45], and VIFNet [42]. Among them, VIFNet and DCMNet are multimodal methods. VIFNet utilizes both RGB and infrared data for joint dehazing, while DCMNet leverages RGB images with corresponding depth maps for haze removal. Each of the methods was retrained on the VTUAV and CART datasets under the same settings as ours, with hyperparameters adopted from the official implementations. The performance of these methods was evaluated using two standard metrics: PSNR and SSIM. Quantitative results are presented in Tab. 2 and Tab. 3, respectively. Qualitative visual results are shown in Fig. 5 and Fig. 6.

(i) On the VTUAV dataset, experimental results demonstrate a clear performance advantage of our method over SoTA methods across all categories, as visualized in Fig. 5. Our proposed method achieves the highest PSNR and SSIM values for each category, with an average PSNR of 31.57 and SSIM of 0.918, substantially outperforming the second-best method, VIFNet, which achieved 29.21 (PSNR) and 0.881 (SSIM). This corresponds to an improvement of 8.1% in PSNR and 4.2% in SSIM. Notably, the improvements are more evident in complex scenarios such as Street and Car, where

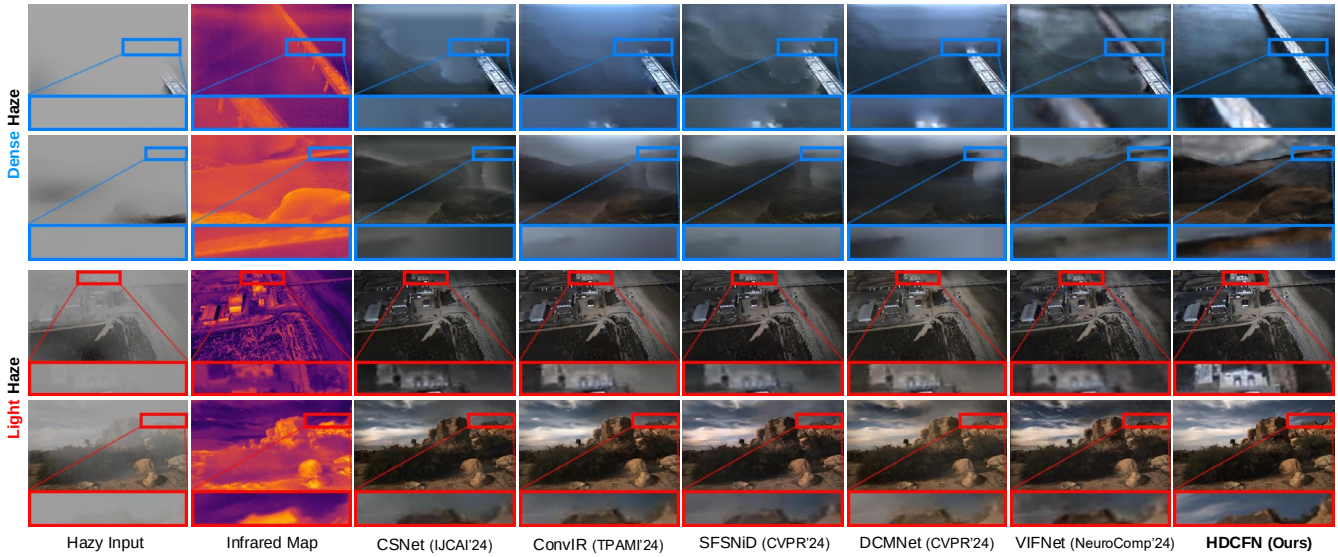

Figure 6: Visualization of partial experimental results on the CART dataset for dense and light haze conditions.

Table 3: Comparison of PSNR/SSIM results on the CART dataset with SoTA dehazing methods (Dense vs. Light haze).

Method	Modality	Bridge		Lake		River		Rock		**Average**	
		Dense	Light	Dense	Light	Dense	Light	Dense	Light	Dense	Light
CSNet	RGB	20.80/0.795	22.74/0.865	20.87/0.821	25.42/0.907	17.32/0.807	22.24/0.882	19.93/0.756	23.68/0.867	19.73/0.795	23.52/0.880
ConvIR	RGB	21.39/0.805	26.15/0.872	21.42/0.831	27.20/0.910	19.85/0.830	25.25/0.891	21.48/0.768	26.65/0.875	21.04/0.808	26.31/0.887
SFSNiD	RGB	21.23/0.802	26.43/0.873	22.13/0.830	27.41/0.908	18.97/0.829	23.31/0.886	20.95/0.765	25.99/0.871	20.82/0.806	25.78/0.884
DCMNet	RGB+Depth	21.43/0.803	26.48/0.881	22.39/0.832	27.82/0.915	21.41/0.842	25.30/0.896	20.76/0.769	27.19/0.875	21.49/0.810	26.69/0.891
VIFNet	RGB+IR	22.12/0.809	26.54/0.881	23.14/0.837	28.25/0.917	21.37/0.833	25.84/0.897	22.79/0.773	28.53/0.879	22.36/0.813	27.29/0.893
Ours	RGB+IR	**23.51/0.834**	**28.15/0.894**	**24.86/0.896**	**29.43/0.935**	**23.78/0.869**	**27.98/0.907**	**24.37/0.837**	**29.40/0.911**	**24.13/0.859**	**28.74/0.912**
Gain (%)	–	6.3%/3.1%	6.1%/1.5%	7.4%/7.0%	4.2%/2.0%	11.0%/3.2%	8.3%/1.1%	6.9%/8.3%	3.0%/3.6%	**7.9%/5.7%**	**5.3%/2.1%**

our method achieves PSNR gains of 8.8% and SSIM gains of 8.4%, respectively. As shown in the first row of Fig. 5, in areas with dense haze occlusion, other methods fail to recover small ground targets (e.g., the blue sedan). In contrast, our method not only restores the object contours but also recovers its color. These results highlight the robustness of our method in reconstructing fine details and preserving structural integrity, even in challenging conditions.

(ii) On the CART dataset, our method consistently surpasses competing approaches, with notable improvements observed in dense haze scenarios, as shown in Fig. 6. In dense haze conditions, our method achieves an average PSNR of 24.13 and SSIM of 0.859, surpassing the second-best method, VIFNet, which attains 22.36 (PSNR) and 0.813 (SSIM). This leads to an improvement of 7.9% in PSNR and 5.7% in SSIM. In light haze conditions, our method also demonstrates superior performance, achieving PSNR and SSIM values of 28.74 and 0.912, respectively, compared to VIFNet's 27.29 (PSNR) and 0.893 (SSIM). These results highlight the effectiveness of our approach, with the notable improvements in dense haze conditions suggesting that the integration of infrared signals enhances the model's ability to capture essential structural and textural details obscured by haze.

5.3 Real-World Evaluation and Analysis

To evaluate the real-world performance of the proposed method, we conducted experiments on the CityUAV dataset, which is captured under foggy weather across various urban scenes. The models were pretrained on the VTUAV and CART datasets and then applied to the CityUAV dataset without further fine-tuning. Qualitative results are presented in Fig. 7, while quantitative results are summarized in Tab. 4 and visualized in Fig. 8.

For the quantitative evaluation, we employed two no-reference image quality metrics: the Natural Image Quality Evaluator (NIQE) [26] and the Perception-based Image Quality Evaluator (PIQE) [35]. NIQE quantifies perceptual quality by measuring statistical deviations from natural scene statistics, offering an objective assessment of image quality. PIQE evaluates local and global image characteristics to estimate perceptual distortions, providing a quantification of perceptual impairments in image structure and content.

Experimental findings can be summarized as follows:

(i) Quantitative Analysis: The proposed HDCFN achieves the best performance on NIQE and PIQE metrics, as detailed in Tab. 4. Specifically, HDCFN outperforms the second-best method with a

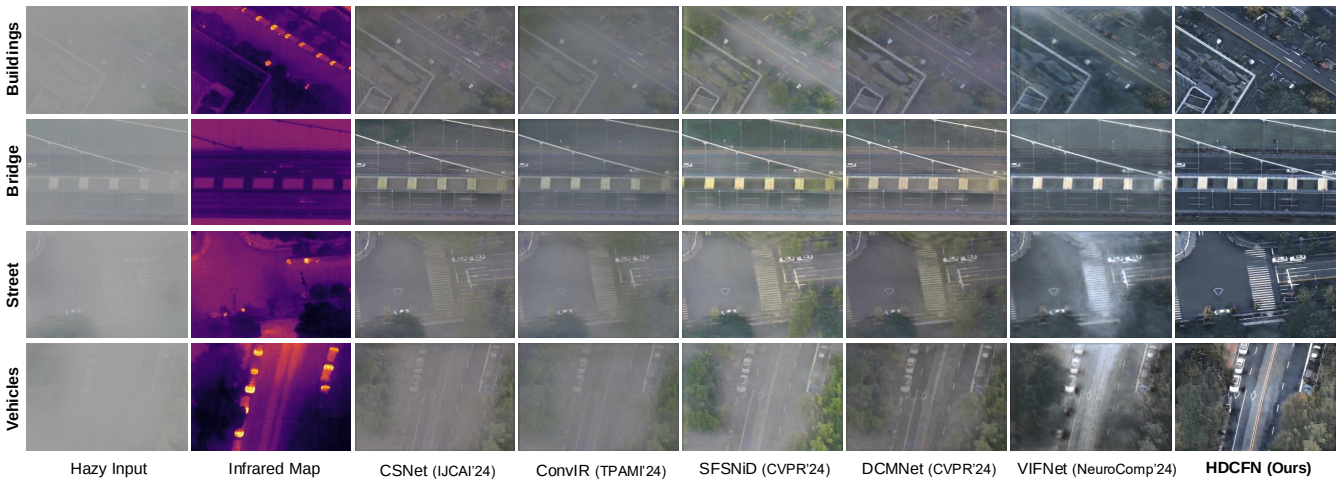

Figure 7: Visualization of representative dehazing results from the CityUAV dataset, captured by a UAV in foggy urban.

Table 4: NIQE and PIQE score comparison across methods.

Metric	CSNet	ConvIR	SFSNiD	DCMNet	VIFNet	**HDCFN**	Gain
NIQE ↓	19.824	18.335	19.640	17.527	16.983	**15.379**	9.4%
PIQE ↓	42.532	39.682	41.168	37.419	36.926	**34.407**	6.8%

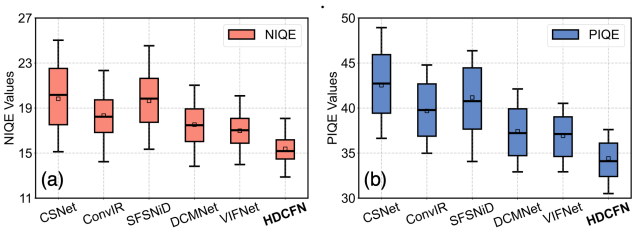

Figure 8: Evaluation results on the CityUAV dataset.

9.4% reduction in NIQE score (15.379 vs. 16.983) and a 6.8% reduction in PIQE score (34.407 vs. 36.926). These results demonstrate that HDCFN produces perceptually superior images by reducing statistical deviations and perceptual distortions.

(ii) Qualitative Analysis: The visual results visualized in Fig. 7 highlight the superior dehazing performance of HDCFN. Compared to other dehazing methods, HDCFN restores fine details such as building textures and small cars more clearly, avoiding the over-smoothing and artifacts observed in competing approaches. Furthermore, HDCFN maintains structural coherence and contrast, effectively mitigating haze artifacts while preserving realistic color tones, resulting in dehazed images that are visually appealing and accurate to the original scene.

(iii) Robustness Analysis: We further evaluated the robustness of different methods. Fig. 8 (a) and (b) show the statistical distributions of NIQE and PIQE scores on the CityUAV dataset. The results indicate that, under dense haze conditions, our method exhibits the lowest variance in both NIQE and PIQE, while methods like VIFNet and DCMNet exhibit higher variability. These findings highlight the robustness of HDCFN in dehazing aerial images and its promising generalization to real-world data for practical UAV applications.

Table 5: Evaluation results on object detection task.

Metric	IGNet	CDDF	EMMA	DCMNet	VIFNet	**HDCFN**	Gain
mAP ↑	0.364	0.417	0.433	0.575	0.607	**0.732**	20.5%
Recall ↑	0.411	0.448	0.472	0.609	0.634	**0.755**	19.1%

5.4 Evaluation on Object Detection Task

We compare HDCFN with SoTA Visible-Infrared Fusion (VIF) methods and dehazing methods on downstream visual tasks, specifically object detection, to investigate whether enhanced image clarity and accuracy can improve performance in high-level vision tasks. The experiments were conducted on the VTUAV dataset, a UAV-captured dataset comprising infrared and visible light images with precise bounding box annotations.

Fig. 9 presents the output results from different VIF methods (IGNet [19], CDDF [47], and EMMA [48]) as well as dehazing methods (DCMNet, VIFNet, and our HDCFN) on the VTUAV dataset. The visual results demonstrate that, while VIF methods recover the general shape of objects obscured by dense haze, they fail to preserve fine details and color accuracy. Other dehazing algorithms, such as VIFNet and DCMNet, struggle with residual haze and blurry contours in areas with heavy haze. In contrast, HDCFN effectively integrates haze-resistant infrared features with visible light, producing sharper boundaries and more natural color tones.

For the object detection task, we applied a pre-trained YOLOv8m model to both VIF and dehazed images without further fine-tuning. As shown in Fig. 9, HDCFN demonstrates superior detection accuracy (mAP) and higher Recall, indicating fewer false positives and missed detections compared to the VIF and other dehazing methods. Quantitative results recorded in Tab. 5 further highlight HDCFN's superiority, leading to improved performance in downstream tasks such as object detection.

6 Ablation Study

The ablation study was conducted on the VTUAV and CART datasets, incorporating five distinct test cases: IR-only (infrared input), RGB-only (RGB input, baseline), and three variations of the proposed

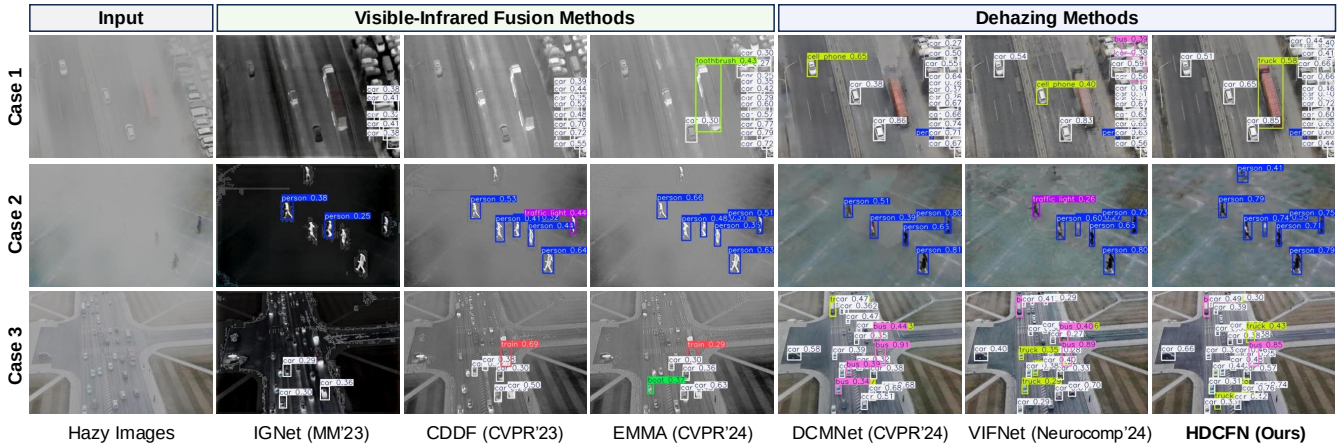

Figure 9: Visualization of object detection results on the outputs generated by VIF methods and dehazing methods.

Table 6: Ablation study results on the VTUAV and CART datasets, evaluating the impact of loss functions and algorithmic modules on dehazing performance.

Case	Loss		Module		VTUAV		CART	
	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{perc}}$	IMFE	CMAF	PSNR	SSIM	PSNR	SSIM
IR-only	✓				23.24	0.819	18.86	0.779
RGB-only	✓				25.47	0.802	20.03	0.764
(A)	✓		✓		28.01	0.863	22.07	0.820
(B)	✓		✓	✓	30.74	0.906	23.35	0.846
(C)	✓	✓	✓	✓	**31.57**	**0.918**	**24.13**	**0.859**

model: (A) baseline with IMFE, (B) IMFE combined with CMAF, and (C) the full model. We evaluated reconstruction fidelity using PSNR and structural similarity with SSIM.

As shown in Tab. 6, each model component contributes incrementally to overall performance. Compared to baseline model, IMFE (Case A) improves PSNR by 2.54 dB on VTUAV dataset and 2.04 dB on CART dataset, and increases SSIM by 0.61 and 0.56, respectively, demonstrating its effectiveness in extracting multi-scale features crucial for recovering fine details. Adding CMAF (Case B) further enhances PSNR by 2.73 dB and SSIM by 0.43 on VTUAV dataset, highlighting its effectiveness in adaptively adjusting features fusion across modalities. The full model (Case C), which integrates the perception loss term $\mathcal{L}_{\text{perc}}$, achieves the highest performance with PSNR values of 31.57 dB on VTUAV dataset and 24.13 dB on CART dataset, emphasizing the contribution of the perception loss in refining visual quality and initial feature alignment. These results validate the cumulative benefits of each component.

Tab. 7 summarizes the performance of the CMAF module in comparison with other feature fusion methods, including element-wise addition (Addition), channel-wise concatenation (Concat), and cross-attention fusion (Cross-Atten.) [46]. On the VTUAV dataset, CMAF outperforms the other methods by 9.5% in PSNR and 5.3% in SSIM. Similarly, on the CART dataset, CMAF achieves a 7.1% improvement in PSNR and 3.9% in SSIM. These results highlight the effectiveness of CMAF in dynamic feature fusion according to haze distribution from both visible and infrared inputs, effectively capturing the complementary information from each modality.

Table 7: Performance comparison of feature fusion methods and the proposed CMAF module.

Dataset	Metric	Addition	Concat	Cross-Atten.	**CMAF**	Gain
VTUAV	PSNR	27.94	28.01	28.83	**31.57**	9.5%
	SSIM	0.859	0.863	0.872	**0.918**	5.3%
CART	PSNR	21.85	22.07	22.54	**24.13**	7.1%
	SSIM	0.816	0.820	0.827	**0.859**	3.9%

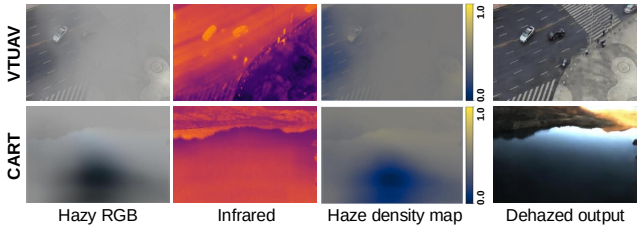

Figure 10: Adaptive weighting of visible and infrared features under various haze densities.

Fig. 10 illustrates the haze density map generated by the Haze Distribution Estimator (HDE) in CMAF. During UAV flight, both the UAV and haze are dynamic, with uneven haze density leading to spatially varying occlusions. Based on the predicted haze density map, CMAF assigns greater weight to infrared features in heavily hazy areas and more weight to visible features in regions with lighter occlusion, as formulated in Eq. (12). This dynamic weighting effectively leverages the complementary strengths of both modalities, enhancing the perceptual quality and preserving fine details.

7 Conclusion

This paper proposes an infrared-guided image dehazing method for UAV scenarios, which exploits the complementary strengths of visible and infrared modalities. This approach integrates a multi-scale feature enhancement framework with a cross-modal adaptive fusion module, facilitating recovery of fine-grained details and dynamic adaptation to varying haze distributions. Extensive experiments demonstrate the method's effectiveness and robustness, achieving SoTA performance, particularly in dense haze scenarios.

References

[1] Tianxiang Chen, Qi Chu, Bin Liu, and Nenghai Yu. 2023. Fluid Dynamics-Inspired Network for Infrared Small Target Detection.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 590–598.

[2] De Cheng, Yan Li, Dingwen Zhang, Nannan Wang, Xinbo Gao, and Jiande Sun. 2021. Robust single image dehazing based on consistent and contrast-assisted reconstruction. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 848–854.

[3] Xiaofeng Cong, Jie Gui, Jing Zhang, Junming Hou, and Hao Shen. 2024. A Semi-supervised Nighttime Dehazing Baseline with Spatial-Frequency Aware and Realistic Brightness Constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2631–2640.

[4] Yuning Cui, Mingyu Liu, Wenqi Ren, and Alois Knoll. 2024. Hybrid Frequency Modulation Network for Image Restoration. In *International Joint Conference on Artificial Intelligence*. 722–730.

[5] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. 2024. Revitalizing Convolutional Network for Image Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 9423–9438.

[6] Fabian Erlenbusch, Constanze Merkt, Bernardo de Oliveira, Alexander Gatter, Friedhelm Schwenker, Ulrich Klauck, and Michael Teutsch. 2023. Thermal infrared single image dehazing and blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 459–469.

[7] Yuxin Feng, Long Ma, Xiaozhe Meng, Fan Zhou, Risheng Liu, and Zhuo Su. 2024. Advancing real-world image dehazing: perspective, modules, and training. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 43, 12 (2024), 9303–9320.

[8] Jie Gui, Xiaofeng Cong, Yuan Cao, Wenqi Ren, Jun Zhang, Jing Zhang, and Dacheng Tao. 2021. A comprehensive survey on image dehazing based on deep learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 4426–4433.

[9] Xiaojie Guo, Yang Yang, Chaoyue Wang, and Jiayi Ma. 2022. Image dehazing via enhancement, restoration, and fusion: A survey. *Information Fusion* 86 (2022), 146–170.

[10] Kaiming He, Jian Sun, and Xiaoou Tang. 2010. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 12 (2010), 2341–2353.

[11] Shaohui Jin, Ziqin Xu, Mingliang Xu, and Hao Liu. 2024. Time-gated imaging through dense fog via physics-driven swin transformer. *Optics Express* 32, 11 (2024), 18812–18830.

[12] Pirunthan Keerthinathan, Narmilan Amarasingam, Grant Hamilton, and Felipe Gonzalez. 2023. Exploring unmanned aerial systems operations in wildfire management: Data types, processing algorithms and navigation. *International Journal of Remote Sensing* 44, 18 (2023), 5628–5685.

[13] Connor Lee, Matthew Anderson, Nikhil Ranganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. 2024. Caltech Aerial RGB-Thermal Dataset in the Wild. In *European Conference on Computer Vision*. 236–256.

[14] Chenyang Li, Suiping Zhou, Hang Yu, Tianxiang Guo, Yuru Guo, and Jichen Gao. 2024. An efficient method for detecting dense and small objects in uav images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2024), 6601–6615.

[15] Hongyu Li, Jia Li, Dong Zhao, and Long Xu. 2021. DehazeFlow: Multi-scale conditional flow network for single image dehazing. In *Proceedings of the ACM International Conference on Multimedia*. 2577–2585.

[16] Huafeng Li, Junyu Liu, Yafei Zhang, and Yu Liu. 2024. A deep learning framework for infrared and visible image fusion without strict registration. *International Journal of Computer Vision* 132, 5 (2024), 1625–1644.

[17] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. 2023. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 11040–11052.

[18] Huafeng Li, Junzhi Zhao, Jinxing Li, Zhengtao Yu, and Guangming Lu. 2023. Feature dynamic alignment and refinement for infrared–visible image fusion: Translation robust fusion. *Information Fusion* 95 (2023), 26–41.

[19] Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. 2023. Learning a Graph Neural Network with Cross Modality Interaction for Image Fusion. In *Proceedings of the ACM International Conference on Multimedia*. 4471–4479.

[20] Yudong Liang, Bin Wang, Wangmeng Zuo, Jiaying Liu, and Wenqi Ren. 2022. Self-supervised Learning and Adaptation for Single Image Dehazing.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1137–1143.

[21] KN Liou. 2002. An Introduction to Atmospheric Radiation. *International Geophysics Series* 84 (2002).

[22] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. 2023. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8115–8124.

[23] Ye Liu, Liang Wan, Huazhu Fu, Jing Qin, and Lei Zhu. 2022. Phase-based memory network for video dehazing. In *Proceedings of the ACM International Conference on Multimedia*. 5427–5435.

[24] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. 2023. Image Restoration with Mean-Reverting Stochastic Differential Equations. In *International Conference on Machine Learning*. 23045–23066.

[25] EJ McCartney. 1976. Optics of the atmosphere: scattering by molecules and particles.

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters* 20, 3 (2012), 209–212.

[27] Srinivasa G Narasimhan and Shree K Nayar. 2002. Vision and the atmosphere. *International Journal of Computer Vision* 48 (2002), 233–254.

[28] Lord Rayleigh. 1871. On the light from the sky, its polarization and colour. *Phil Mag* 41 (1871), 274.

[29] Minxian Shen, Tianyi Lv, Yi Liu, Jialiang Zhang, and Mingye Ju. 2024. A Comprehensive Review of Traditional and Deep-Learning-Based Defogging Algorithms. *Electronics* 13, 17 (2024), 3392.

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

[31] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the ACM International Conference on Multimedia*. 4003–4011.

[32] Wei Tang, Fazhi He, and Yu Liu. 2024. ITFuse: An interactive transformer for infrared and visible image fusion. *Pattern Recognition* 156 (2024), 110822.

[33] Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. 2023. DATFuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 7 (2023), 3159–3172.

[34] Kashif Usmani, Timothy O'Connor, Pranav Wani, and Bahram Javidi. 2022. 3D object detection through fog and occlusion: passive integral imaging vs active (LiDAR) sensing. *Optics Express* 31, 1 (2022), 479–491.

[35] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. 2015. Blind image quality evaluation using perception based features. In *National Conference on Communications*. 1–6.

[36] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. 2022. Unsupervised Misaligned Infrared and Visible Image Fusion via Cross-Modality Image Generation and Registration. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 3508–3515.

[37] Hebaixu Wang, Hao Zhang, Xunpeng Yi, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. 2024. TeRF: Text-driven and region-aware flexible visible and infrared image fusion. In *Proceedings of the ACM International Conference on Multimedia*. 935–944.

[38] Yangyang Wang, Jie Zhang, and Jian Zhou. 2024. Urban traffic tiny object detection via attention and multi-scale feature driven in UAV-vision. *Scientific Reports* 14, 1 (2024), 20614.

[39] Gang Wu, Junjun Jiang, Kui Jiang, and Xianming Liu. 2024. Harmony in diversity: Improving all-in-one image restoration via multi-task collaboration. In *Proceedings of the ACM International Conference on Multimedia*. 6015–6023.

[40] Jie Xiao, Xueyang Fu, Man Zhou, Hongjian Liu, and Zheng-Jun Zha. 2023. Random shuffle transformer for image restoration. In *International Conference on Machine Learning*. 38039–38058.

[41] Tianyang Xu, Yifan Pan, Zhenhua Feng, Xuefeng Zhu, Chunyang Cheng, Xiao-Jun Wu, and Josef Kittler. 2024. Learning Feature Restoration Transformer for Robust Dehazing Visual Object Tracking. *International Journal of Computer Vision* (2024), 1–18.

[42] Meng Yu, Te Cui, Haoyang Lu, and Yufeng Yue. 2024. VIFNet: An end-to-end visible-infrared fusion network for image dehazing. *Neurocomputing* (2024), 128105.

[43] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 8886–8895.

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

[45] Yafei Zhang, Shen Zhou, and Huafeng Li. 2024. Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 2846–2855.

[46] Junwei Zhao, Shiliang Zhang, Zhaofei Yu, and Tiejun Huang. 2024. Recognizing ultra-high-speed moving objects with bio-inspired spike camera. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7478–7486.

[47] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. 2023. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5906–5916.

[48] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. 2024. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25912–25921.

[49] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. 2023. Fourmer: An efficient global modeling paradigm for image restoration. In *International Conference on Machine Learning*. 42589–42601.