Pattern Recognition 171 (2026) 112206

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Image dehazing via RGB-FIR multimodal fusion and collaborative learning

Ruolin Du [a,1], Han Wang [a,1], Wenjie Liu [a], Guangcheng Wang [a,*], Kui Jiang [b], Hanseok Ko [c]

[a] School of Transportation and Civil Engineering, Nantong University, Nantong, 226019, China
[b] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
[c] School of Electrical Engineering, Korea University, Seoul, 02841, South Korea

ARTICLE INFO

Keywords:
Image dehazing
RGB-FIR multimodal
Multi-level feature fusion
Multi-scale image transfer
Collaborative optimization

ABSTRACT

The mainstream deep learning-based image dehazing methods that rely solely on RGB information experience a sharp decline in performance when confronted with dense fog. To this aim, this paper proposes an RGB-FIR multimodal fusion-transfer network, along with a collaborative optimization strategy tailored for image dehazing. The developed network comprises two key modules: a multi-level feature fusion module and a multi-scale image transfer module. Specifically, the former employs a cross-modal multi-scale based on cross-pooling attention model to explore the complementary information cross scales and levels of RGB-FIR multimodal images. The latter characterizes the mapping relationship between the fused multimodal features and the ground-truth image. To further boost the optimization and facilitate the dehazing performance, a collaborative optimization strategy is elaborated to harmonies the merits of both generative adversarial loss, feature matching loss, perceptual loss, and fidelity loss. To validate the effectiveness of our proposed dehazing algorithm, we have collected 11,736 pairs of foggy and foggy-free images using a binocular RGB-FIR camera. Both subjective and objective experiments demonstrate that our RGB-FIR multimodal dehazing algorithm outperforms existing state-of-the-art (SOTA) image dehazing methods in terms of restoring details, textures, and color information from foggy images.

1. Introduction

On foggy days, water droplets suspended in the air cause severe atmospheric light scattering due to oversaturation and high density, drastically reducing the image quality captured by cameras [1]. The decline in image quality caused by foggy days leads to difficulties in Object detection and recognition in artificial intelligence systems and presents new challenges for feature extraction in deep learning models. Thus, image dehazing has become a hot research topic in the field of computer vision [2]. Existing image dehazing methods are classified into the prior knowledge-based dehazing methods, the deep learning-based dehazing methods according to the dehazing principles. The prior knowledge-based dehazing methods utilize specific prior knowledge to generate haze-free images. The prior knowledge encapsulates the inherent objective rules of specific attributes present in digital images captured in foggy environments. Attributes such as the dark channel prior and color attenuation prior serve as crucial indicators, enabling the deduction of the intricate mapping relationship between hazy and clean images [3,4].

The deep learning-based dehazing methods mainly rely on Convolutional Neural Networks (CNNs) to generate haze-free images. Its core lies in utilizing CNNs to accurately estimate the transmission rate and atmospheric light, and then restoring the image based on the physical atmospheric scattering model (ASM). With the rise of Generative Adversarial Networks (GANs) [5], a series of GANs-based image dehazing studies have developed rapidly. Based on generative adversarial learning, they have broken the limitations of the physical atmospheric scattering model, making the image dehazing process more flexible and efficient.

Early dehazing methods were based on single RGB images and utilized prior knowledge and ASM to effectively restore background details and color information in foggy images, achieving dehazing. However, light cannot penetrate in dense fog conditions, both RGB and NIR images tend to lose almost all texture and detail features. Consequently, the two types of dehazing methods mentioned earlier still struggle to recover high-definition, fog-free images when confronted with dense fog. The scattering characteristics and imaging features of light vary depending on its wavelength. Fig. 1 presents a set of examples demonstrating image dehazing perfrmance based on NIR and FIR in dense fog conditions. Among them, Fig. 1(a)–(c) represent RGB, NIR, and FIR images in foggy weather, respectively, while Fig. 1(f) shows their corresponding clean

* Corresponding author.
E-mail addresses: 2233320008@stmail.ntu.edu.cn (R. Du), hanwang@ntu.edu.cn (H. Wang), lwj2014@ntu.edu.cn (W. Liu), wanggc@ntu.edu.cn (G. Wang), jiangkui@hit.edu.cn (K. Jiang), hsko@korea.ac.kr (H. Ko).
[1] Co-first authors.

https://doi.org/10.1016/j.patcog.2025.112206
Received 29 May 2024; Received in revised form 15 May 2025; Accepted 22 July 2025
Available online 26 July 2025

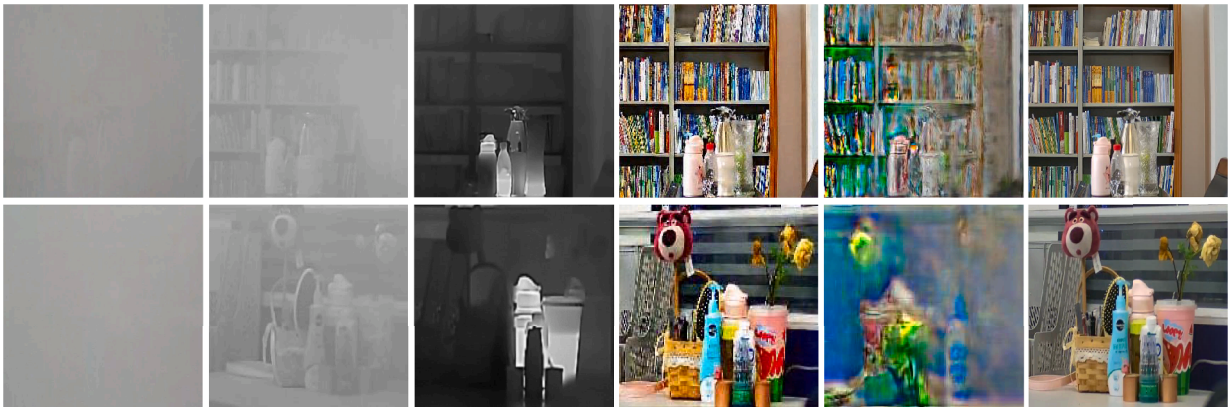


Fig. 1. Performance comparison of image dehazing based on RGB-NIR fusion and RGB-FIR fusion in dense fog weather.

image. Fig. 1(d) and (e) depict the dehazing results based on RGB-NIR multimodal fusion and RGB-FIR multimodal fusion, respectively. RGB spans a wavelength range of 380 nanometers to 750 nanometers, which can exhibit rich colors and delicate textures in images under sufficient illumination. NIR, with a wavelength range of 760 nanometers to 1500 nanometers, can capture clearer texture information when combined with a dedicated NIR filter, a photosensitive sensor, and a lens. This information is presented in grayscale images, making it crucial for applications such as night vision surveillance. FIR has a wavelength range of 3000 nanometers to 1 mm . The Rayleigh scattering model [6] reveals an important physical principle: the intensity of scattered light in the atmosphere is inversely proportional to the wavelength of the incident light. This means that longer wavelengths of light have stronger abilities to penetrate hazy atmospheres. Therefore, combining NIR or FIR with multimodal dehazing methods is an feasible solution. Based on this principle, existing image fusion-based dehazing models primarily focus on the processing of RGB-NIR multispectral images. These models cleverly adopt a two-stage dehazing strategy of "fusion first, color enhancement later" to generate haze-free images. Specifically, they first fuse RGB and NIR images from two different modalities to endow the image with the ability to "penetrate the haze". Subsequently, color enhancement is performed on the fused grayscale image to address potential color distortion issues that may arise during the dehazing process.

By comparison, it is evident that the wavelength of FIR is significantly larger than that of visible light and near-infrared light, making FIR images exhibit stronger "haze transparency" under dense fog conditions, as shown in Fig. 1(c). To address this challenge, this paper proposes an end-to-end RGB-FIR multimodal fusion-transfer dehazing network and its collaborative optimization strategy for image dehazing. The main contributions of our work are as follows:

(1) We propose an end-to-end RGB-FIR multimodal fusion-transfer dehazing network framework. Compared to state-of-the-art (SOTA) image dehazing methods, on our self-build indoor dense fog dataset, PSNR improved by 1.9 dB, SSIM increased by 6 %, and MSE decreased by 420.7. On the outdoor dense fog dataset, PSNR improved by 0.52 dB, SSIM increased by 2 %, and MSE decreased by 170.9.
(2) We develop a generative adversarial-based collaborative optimization strategy which adaptively integrate various losses to update the dehazing network parameters.
(3) We present a novel multi-scale cross-modal attention module that efficiently integrates pixels from RGB-FIR multimodal image pairs across diverse scales through the utilization of cross-pooling technique.
(4) In a wild environment, we collected 11.736 k pairs of RGB-FIR multimodal foggy and foggy-free images, which were used for performance verification of the proposed dehazing method and related SOTA algorithms.

2. Related works

2.1. Prior knowledge-based image dehazing

In the early stages of image dehazing research, the focus was primarily on leveraging prior knowledge. He et al. revolutionized the field with their Dark Channel Prior (DCP), which capitalized on the observation of extremely low pixel values in certain channels within non-sky regions [4]. This discovery led to the development of an efficient dehazing model. Zhu et al. further expanded the theoretical framework by introducing the Color Attenuation Prior (CAP), which explored the proportional relationship between haze concentration and image brightness and saturation [7]. Their work culminated in the creation of a linear model that integrated these factors, providing deeper insights into the dehazing process. Building upon the DCP, Yadav et al. made significant contributions by incorporating the Non-local Haze Line Averaging technique [8]. Furthermore, they introduced the Robust Multiscale Weighting-based Edge-Smoothing Filter, a sophisticated tool that fine-tuned and optimized the transmittance map, enabling the precise restoration of haze-free images.

2.2. Deep learning-based image dehazing

Deep learning has significantly advanced image dehazing, primarily leveraging Convolutional Neural Networks (CNNs) to craft efficient dehazing models. Cai et al. pioneered DehazeNet, which estimated medium transmission rates and restored clean images using the atmospheric scattering model [9]. However, its reliance on a two-stage process rooted in physical models limited its optimization potential despite CNNs' superiority in transmission rate estimation. To push the boundaries, Zhang et al. designed DCPDN, a densely connected pyramid network, capable of jointly learning transmission rates, atmospheric light, and the dehazing process, thus bridging the gap between physical models and deep learning techniques [10]. Guo et al. combined the strengths of Transformer and CNN, leveraging the Transformer's global contextual perception and the CNN's local detail processing abilities to further enhance dehazing outcomes [11]. Wang et al. proposed the defogging network USCFormer [12], which restores hazy images through contrast constraints while preserving the colors and details of the defogged images by incorporating semantic priors and intra-object semantic associations.

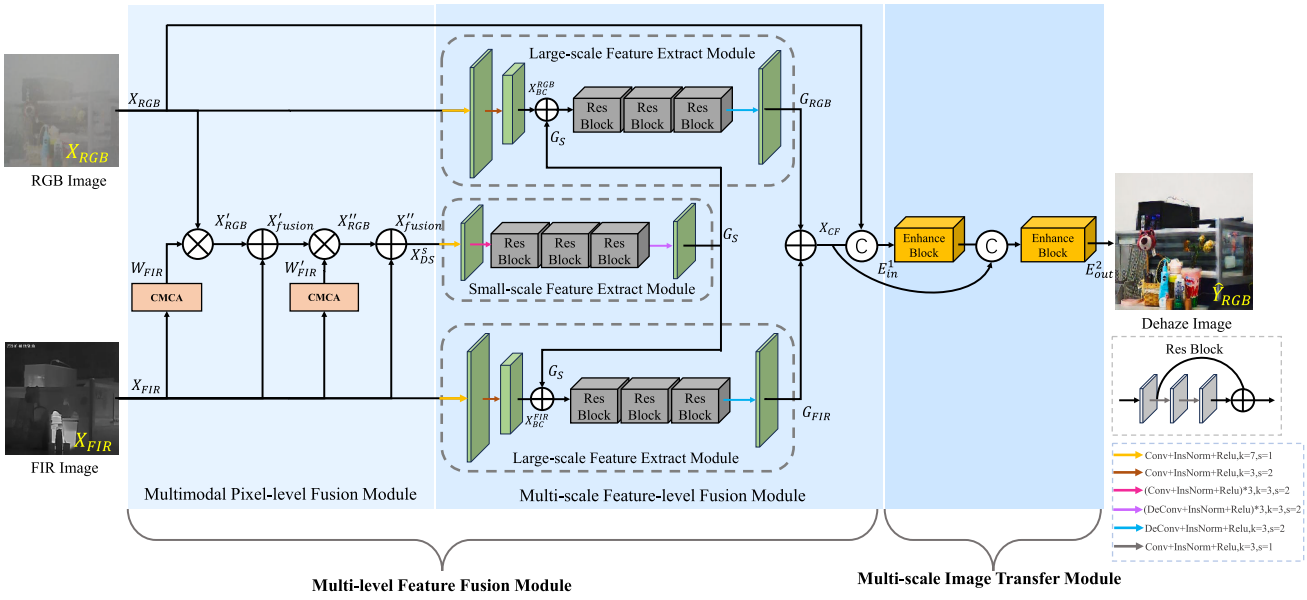


Fig. 2. Pipeline of the RGB-FIR multimodal fusion-transfer network.

With the rise of Generative Adversarial Networks (GANs), researchers have begun to explore their potential in image dehazing tasks. Such as, Qu et al. proposed the multi-scale image transformation-based EPDN dehazing network, which integrates a resolution generator, a multi-scale discriminators, and an enhancer [13]. Through the supervision of the discriminator on the generator, it produces coarse-grained images, which are then further refined by the enhancer. However, GANs often rely on paired training datasets in dehazing tasks, limiting their application scope to a certain extent. To address this challenge, Engin et al. introduced the Cycle-Dehaze network [14] architecture, which incorporates cyclic perception consistency loss to effectively dehaze images even in the absence of paired training data. In order to improve the generalization ability of the defogging network to real hazy images, Wang et al. proposed an unsupervised Cycle-SNSPGAN [15], which utilizes an SN-Soft-Patch GAN and a new cycle self-perceiving loss and color loss to bring the defogged image to the desired brightness.

2.3. *Image fusion-based image dehazing*

The image fusion-based dehazing studies aim to design dehazing algorithms by combining multi-spectral images with different dehazing capabilities. RGB images possess rich color and texture features, but due to the characteristics of visible light wavelengths, they are not conducive to image dehazing tasks under haze weather. Therefore, existing multi-spectral image fusion dehazing algorithms typically adopt a strategy of fusion followed by color enhancement. For instance, Shibata et al. proposed a fusion algorithm based on local contrast measurement, which effectively extracts salient information by fusing the gradient information of NIR and RGB images [16]. They then employ the Poisson image editing technique to construct brightness information from the fused gradients. Kumar et al. estimated the transmission map based on RGB images and further refined it by fusing NIR image information [17]. Finally, they used the improved transmission map to restore the hazy image. In addition, the fusion of visible and infrared images is developing rapidly [18,19]. This rapid progress not only establishes a solid foundation for advancing multimodal image defogging research, but also demonstrates promising applications across diverse domains including semantic segmentation [20], medical imaging [21,22], and object detection [23]. For example, Liu et al. proposed a task-guided, implicit-searched and meta-initialized deep model to address the image fusion problem in a challenging real-world scenario [24]. It integrates information from downstream tasks to guide the unsupervised learning process of image fusion. Furthermore, it efficiently auto-discovers compact architectures of fusion models by designing an implicit search scheme. Wang et al. developed a framework to improve misaligned multimodal image fusion [25]. This framework achieves coarse-to-fine image alignment through a single-stage optimization process. Additionally, it incorporates a Transformer-Conv based fusion sub-network to perform high-quality fusion of the aligned images.

3. Proposed method

The developed RGB-FIR multimodal fusion-transfer network is comprised of two pivotal components: a multi-level feature fusion module and a multi-scale image transfer module. As illustrated in Fig. 2, the multi-level feature fusion module encompasses two sub-modules: a multimodal pixel-level fusion and a multi-scale feature-level fusion. Through the seamless integration of pixel-level and feature-level fusion techniques, this module achieves modal complementarity, effectively restores intricate image features in dense fog scenarios, and significantly enhances the dehazing effect. This approach addresses the performance limitations of traditional single-modal dehazing algorithms in dense fog environments. Concurrently, the multi-scale image transfer module focuses on learning the intricate nonlinear mapping between the multimodal fusion features observed under dense fog conditions and the corresponding ground-truth images. Finally, by integrating generative adversarial loss, feature matching loss, perceptual loss, and fidelity loss, the RGB-FIR multimodal dehazing network is collaboratively optimized to achieve even more impressive dehazing performance.

3.1. *Multimodal pixel-level fusion module*

To enhance image dehazing performance in dense foggy environments, this paper proposes a multimodal pixel-level fusion module combined with a cross-modal attention mechanism. This approach leverages the haze-penetrating capability of FIR (Far-Infrared) images, effectively integrates multimodal information, and restores degraded textures and details. The pixel-level fusion module employs a three-branch forward fusion strategy, accepting RGB-FIR raw features as its

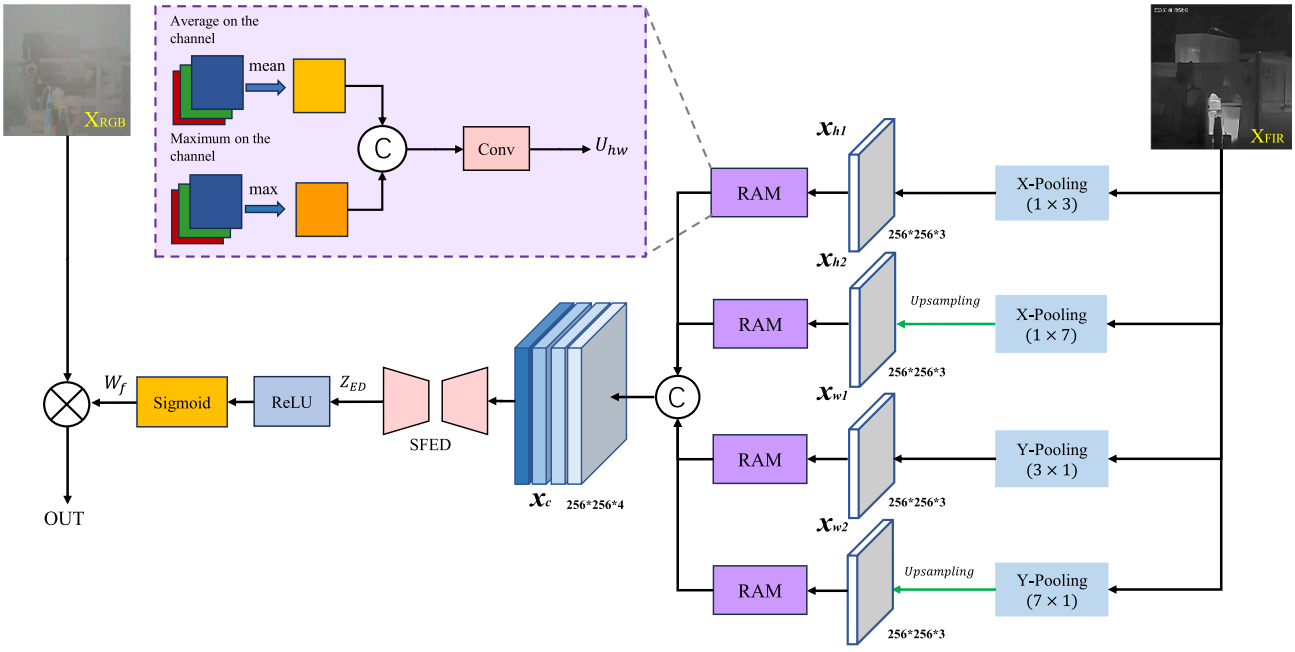


Fig. 3. Structure of the cross-modal multi-scale based on cross-pooling attention (CMCA) module.

input. Precisely, the upper branch concentrates on refining the original RGB pixel features. The lower branch extracts the original FIR pixel features. The intermediary branch integrates the RGB and FIR pixel features to generate a comprehensive RGB-FIR pixel fusion feature stream. To address the lack of texture features in RGB images with dense fog in the underlying network, this paper proposes a novel Cross-modal Multi-scale based on Cross-pooling Attention (CMCA) module, specifically designed for RGB-FIR multimodal pixel-level fusion. This model leverages the fog-penetrating capabilities of thermal radiation in FIR images (pixel values) to restore the missing textures in RGB images affected by dense fog. The steps for fusing RGB-FIR feature streams are as follows: First, the RGB feature map X_{RGB} is adjusted using the weight values W_{FIR} generated by the CMCA module. The optimized X'_{RGB} is then added to the original FIR feature map X_{FIR} to obtain the fused feature X'_{fusion}. Subsequently, the new weight values W'_{FIR} are generated by the CMCA module for a second adjustment, resulting in the feature X''_{RGB}. This X''_{RGB} is further added to the original FIR feature map X_{FIR} for a second time, ultimately yielding the pixel-fused feature X''_{fusion}.

Fig. 3 presents the structure of the proposed CMCA module. The CMCA module takes the original FIR features as input and employs pooling and convolution operations at different scales in both horizontal and vertical cross-directions to extract four streams of low-level features X_{h1}, X_{h2}, X_{w1}, X_{w2}. By using the Representation of Average and Maximum (RAM) module, these feature streams of each branch can be fused for the first time on the channel-level. Then, the feature X_C with cross direction and multi-scale is obtained by a concatenation operation. The feature X_C is then compressed through a spatial feature encoder-decoder (SFED) framework, efficiently condensing the feature channels to derive precise feature mapping values Z_{ED} for each spatial location. Finally, the spatial attention weights W_f are generated using *ReLU* and *Sigmoid* activation functions.

Here, we introduce the detailed definition of the CMCA module. X_{RGB} and X_{FIR} are RGB and FIR feature maps, respectively, $X_{RGB} \in R^{H_i \times W_j \times C_n}$, $X_{FIR} \in R^{H_i \times W_j \times C_n}$. H and W are the length and width of RGB and FIR feature maps, respectively. C is the channel number of RGB and FIR feature maps. First, we conduct two-scale average pooling on the far-infrared feature map in both the horizontal and vertical directions. The specific operations are as follows:

$$X_{h_m} = \frac{1}{k_m} * \sum_{j=1}^{W} f(i,j,n), m \in [1,2], k_m \in [3,7], \tag{1}$$

$$X_{w_m} = \frac{1}{k_m} * \sum_{i=1}^{H} f(i,j,n), m \in [1,2], k_m \in [3,7]. \tag{2}$$

After upsampling adjustment, the spatial resolutions of the four pooling features X_{h1}, X_{h2}, X_{w1}, X_{w2} in the horizontal and vertical directions are aligned. Then, a RAM module is developed to compress the channels of the above four pooling features. The specific calculation process is represented by Eqs. (3)–(6).

$$X_{mean} = Average\left[X_{h_m}^{c_1} + X_{h_m}^{c_2} + \cdots + X_{h_m}^{c_n}\right], \tag{3}$$

$$X_{max} = Max\left[X_{h_m}^{c_1}, X_{h_m}^{c_2}, \ldots, X_{h_m}^{c_n}\right], \tag{4}$$

$$X_{hw} = Concat[X_{mean}, X_{max}], \tag{5}$$

$$U_{hw} = Conv(X_{hw}), \tag{6}$$

where X_{mean} is the pixel average value of all channels. $X_{\max}$ is the maximum value among the channel for a given pixel. *Conv* is a convolution with a kernel size of 1×1. *Concat* is the concatenation operation of the feature channel. Furthermore, the compressed features are concatenated across their channels to derive the integrated feature X_C resulting from four-channel multi-scale cross-pooling.

To extract the significance weights of feature points at different spatial positions from X_c, we develop a spatial feature encoder-decoder (SFED) as shown in Fig. 4. SFED is capable of achieving nonlinear fusion mapping of feature points at the same spatial position across different channels. The specific definition is as follows:

$$\begin{cases} Z_{ED} = F_{ED}(X_c), \\ F_{ED}(x_c) = F_{ex}\{F_{fc2}[\sigma(F_{fc1}(X_c))]\}, \end{cases} \tag{7}$$

where F_{ED} represents a nonlinear fusion mapping function. It comprises a fully-connected dimensionality reduction module, an activation function *ReLU*, and a fully-connected dimensionality enhancement module, as depicted in Fig. 4. F_{ex} represents a reshape operation that maps a one-dimensional optimized feature vector to a two-dimensional spatial feature map, ensuring that the fused feature map has the same dimensions

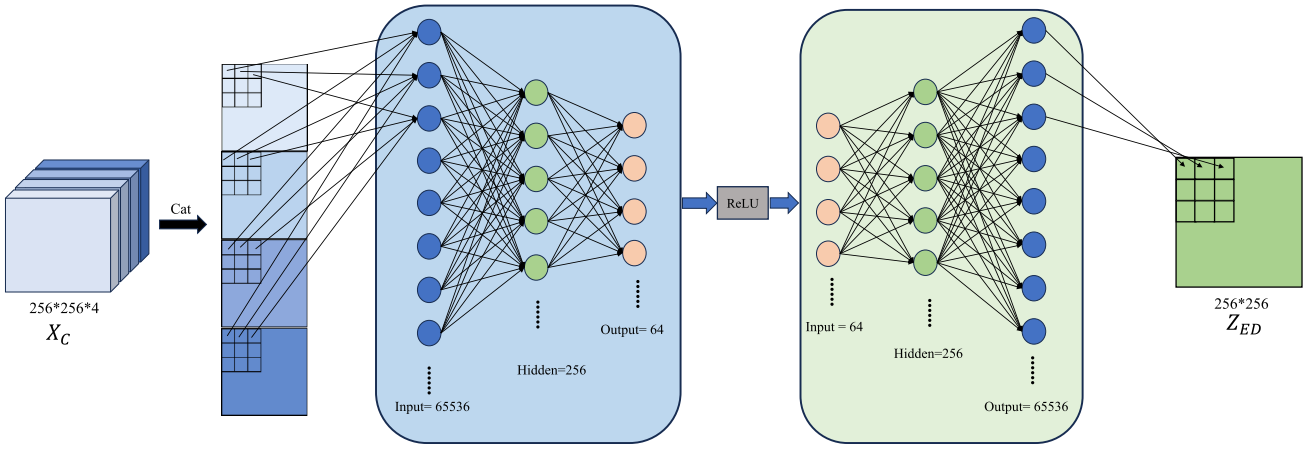


Fig. 4. Spatial feature encoder-decoder (SFED).

as the RGB feature map. F_{fc1} and F_{fc2} are the processes of dimensionality reduction and enhancement processes based on full connections, respectively. Finally, the spatial attention significance weights W_f are derived consecutively using the *ReLU* and *Sigmoid* activation functions, as indicated in the following equation:

$$W_f = \varphi\{\sigma[Conv(Z_{ED})]\}, \tag{8}$$

where *Conv* is a convolution operation with a kernel size of 1x1. σ refers to the *ReLU* activation function. φ refers to the *Sigmoid* activation function. $W_f \in R^{H\times W\times C_n}$.

3.2. Multi-scale feature-level fusion module

To enhance deep-level feature fusion across different modal images and fully leverage the advantages of multi-scale structures in image generation tasks, we design a multi-scale feature-level fusion module. Its architecture is inspired by the multi-scale framework commonly used in image-to-image translation tasks, which progressively generates images from coarse to fine, thereby improving the restoration of structural and detailed information. As shown in the middle of Fig. 2, the multi-scale feature-level fusion module consists of a three-branch generator with two scales. The small-scale feature extraction module is located between two large-scale feature extraction modules, forming a coarse-to-fine and local-to-global multi-scale nested structure. Each branch of the generator is composed of convolution, residual blocks, and deconvolution. The small-scale middle branch first downsamples the multi-modal pixel-level fused feature X''_{fusion} by a factor of two, utilizing residual blocks to extract high-level features. These high-level features are then element-wise summed with the convolutional features from the initial stages of the large-scale upper and lower branches. The deep semantic consistency between modalities is effectively enhanced and the fusion expressiveness of the final generated image is improved. The combined features are further processed in the residual blocks of the large-scale branches, respectively. Finally, the output features from the two large-scale branches are added together to obtain the output result X_{CF} of the multi-level feature fusion network. Then, X_{CF} is concatenated with the original RGB input features in the channel dimension to enhance the color information from the channel dimension, which further improves the color saturation and visual quality of the defogged image. The detailed definition of the small-scale branch is as follows:

$$X^S_{DS} = Downsampling\left(X''_{fusion}\right), \tag{9}$$

$$U_{DS} = US\left\{3 * RB\left[\sigma\left(LN\left(Conv_7\left(X^S_{DS}\right)\right)\right)\right]\right\}, \tag{10}$$

$$G_S = F_{rct}(U_{DS}), \tag{11}$$

where $X''_{fusion} \in R^{H\times W\times C_n}$ is the output of the multimodal pixel-level fusion module. *Downsampling* $(\cdot)$ is a $2\times downsampling$ operation. $Conv_7$ is a convolution with the kernel size of 7 and the step size of 1. σ refers to the *ReLU* activation function. *LN* is the layer normalization. *RB* represents three concatenated residual blocks as shown in the gray part of Fig. 2. *US* represents three deconvolution operations with a kernel size of 3 and a step size of 2, one layer normalization, and one *ReLU* activation function operation. $F_{rct}(\cdot)$ performs the following operations on the feature U_{DS}. First, it applies a 3-valued peripheral pixel-symmetric mirror padding to U_{DS}. Then, it convolves the padded feature with a kernel size of 7. Finally, it activates the convolved result using the *Tanh* function. G_S is the output feature of the small-scale branch.

The inputs of the upper and lower branches are X_{FIR} and X_{RGB}, and the detailed process of their feature fusion is shown in Eqs. (12)–(17).

$$X^{RGB}_{BC} = Conv_3\{\sigma[Conv_7(X_{RGB})]\}, \tag{12}$$

$$X^{FIR}_{BC} = Conv_3\{\sigma[Conv_7(X_{FIR})]\}, \tag{13}$$

$$U_{RGB} = DC\left[3 * RB\left(X^{RGB}_{BC} + G_S\right)\right], \tag{14}$$

$$U_{FIR} = DC\left[3 * RB\left(X^{FIR}_{BC} + G_S\right)\right], \tag{15}$$

$$G_{RGB} = F_{rct}(U_{RGB}), \tag{16}$$

$$G_{FIR} = F_{rct}(U_{FIR}), \tag{17}$$

where $Conv_3$ is a convolution operation with a kernel size of 3 and a step stride of 2. $Conv_7$ is a convolution operation with a kernel size of 7 and a step size of 1. σ refers to the *ReLU* activation function. X^{RGB}_{BC} and X^{FIR}_{BC} are convolutional features obtained through two convolution operations and a *ReLU* activation. $DC(\cdot)$ integrates a deconvolution operation with a kernel size of 3 and a stride of 2, a layer normalization operation, and a ReLU activation operation. U_{RGB} and U_{FIR} are the features obtained from X^{RGB}_{BC} and X^{FIR}_{BC} through the $DC(\cdot)$ operation. G_{RGB} and G_{FIR} are the output features of the upper and lower branches, respectively. After element-wise addition of G_{RGB} and G_{FIR}, they are concatenated with the original RGB image features to produce the feature-level fusion result E^1_{in}. The specific computation steps are outlined in Eqs. (18) and (19).

$$X_{CF} = G_{RGB} + G_{FIR}, \tag{18}$$

$$E^1_{in} = Concat(X_{CF} + X_{RGB}). \tag{19}$$

3.3. Multi-scale image transfer module

The multi-level feature fusion module comprehensively integrates pixel-level and feature-level multimodal information. However, the defogged images may still lack background details or exhibit oversaturation. To address this issue, we propose a pyramid-structured enhancement block designed to refine image details and color information across multiple scales, as illustrated in Fig. 5. After the initial enhancement of the fused feature E^1_{in}, it is concatenated with the intermediate

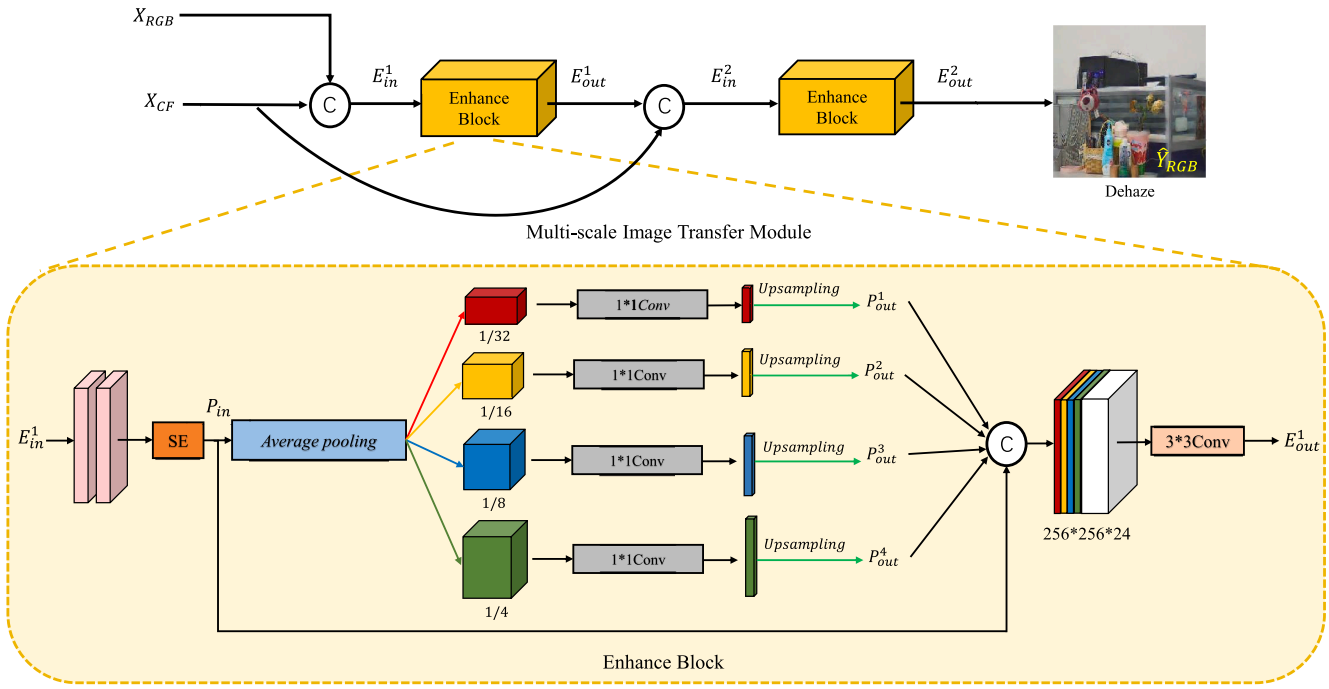


Fig. 5. The structure of multi-scale image transfer module.

output X_{CF} from the multi-level feature fusion module, followed by a second enhancement to yield the final defogged image.

As shown in the yellow box of Fig. 5, the structure of the enhance block employs a multi-scale pyramid design. Firstly, the input feature E^1_{in} is optimized through a channel attention mechanism to emphasize key channel information. Subsequently, the feature undergoes downsampling at four different scales (i.e., 1/4, 1/8, 1/16, and 1/32), generating four sub-feature maps of different resolutions, thus forming a four-branch multi-scale pyramid structure. Then, to fuse these multi-scale features, we upsample each sub-feature map to match the spatial resolution of the original input feature E^1_{in} (after channel attention optimization). Subsequently, the feature maps are merged through channel concatenation, achieving cross-scale information fusion. Finally, a 3×3 convolution is utilized to further integrate the fused features. The specific computational process is as follows:

$$P_{in} = SE\{Conv_3[Conv_3(E^1_{in})]\}, \tag{20}$$

$$P^n_{out} = \begin{cases} Upsampling\{\sigma[Conv(F_{AP}(P_{in},32))]\} & \text{if } n=1, \\ Upsampling\{\sigma[Conv(F_{AP}(P_{in},16))]\} & \text{if } n=2, \\ Upsampling\{\sigma[Conv(F_{AP}(P_{in},8))]\} & \text{if } n=3, \\ Upsampling\{\sigma[Conv(F_{AP}(P_{in},4))]\} & \text{if } n=4, \end{cases} \tag{21}$$

$$E^1_{out} = tanh\{Conv_3[Concat(P^1_{out}, P^2_{out}, P^3_{out}, P^4_{out}, P_{in})]\}, \tag{22}$$

$$E^2_{out} = Conv(E^1_{out} + X_{CF}), \tag{23}$$

$$\hat{Y}_{RGB} = E^2_{out}, \tag{24}$$

where $Conv_3$ is a convolution operation with a kernel size of 3 and the step stride of 1. P_{in} is the input of the pyramid structure. P^n_{out} is the output of the pyramid structure. $F_{AP}(x,b)$ represents performing average pooling on x at scale b to achieve downsampling. σ refers to the *ReLU* activation function. $Upsampling$ is the upsampling function. *Conv* is a convolution operation with a kernel size of 1 and a step size of 1. *tanh* is the activation function *Tanh*. E^1_{out} and E^2_{out} are the features after the first and second enhancements, respectively. $\hat{Y}_{RGB}$ is the final reconstructed hazy-free image.

3.4. Collaborative optimization

For the optimization of the proposed dehazing network, we have designed a collaborative optimization scheme that integrates the generative adversarial loss L_A, the feature matching loss L_{FM}, the perceptual loss L_{VGG}, and the fidelity loss L_F. Fig. 6 shows a visualization of this collaborative optimization scheme. Specifically, we achieve the collaborative optimization of the designed dehazing network by assigning weights to the four types of losses. The detailed implementation process is described as follows:

$$L_{MF} = L_A + \lambda_1 L_{FM} + \lambda_2 L_{VGG} + \lambda_3 L_F, \tag{25}$$

where L_{MF} is the total loss. For the settings of parameters λ_1, λ_2, and λ_3, the detailed discussions are provided in Section 4.2.1.

In the L_A loss, the multi-scale discriminators $D_K(k\in[1,2])$ discriminate the authenticity of the multi-level fused features X_{CF} against the ground-truth image Y_{RGB} and the dehazed image $\hat{Y}_{RGB}$ at different resolutions. The specific definition is as follows:

$$L_A = \min_{\tilde{G}}\Big[\max_{D_K}\sum_{K=1,2}\ell_A(\tilde{G}, D_K)\Big], \tag{26}$$

$$\begin{aligned}\ell_A(\tilde{G}, D_K) = {} & E_{X_{CF},Y_{RGB}}[\log D_K(X_{CF}, Y_{RGB})] \\ & + E_{X_{CF},Y_{RGB},X_{FIR}}\{\log[1 - D_K(X_{CF}, \tilde{G}(X_{RGB}, X_{FIR}))]\},\end{aligned} \tag{27}$$

where $\ell_A(\tilde{G}, D_K)$ is the single adversarial loss of the kth discriminator. $\tilde{G}(X_{RGB}, X_{FIR})$ is the result generated by the multi-scale generator. $E_{X_{CF},Y_{RGB}}$ and $E_{X_{CF},Y_{RGB},X_{FIR}}$ represent the expectation of the corresponding samples. The feature matching loss L_{FM} focuses on the matching of local features, utilizing multi-scale discriminators $D_K(k\in[1,2])$ to discriminate the authenticity of the generated image by the generator $\tilde{G}(X_{RGB}, X_{FIR})$ against the ground-truth image Y_{RGB} at different resolutions. Its specific definition is as follows:

$$L_{FM} = \min_{\tilde{G}}\left[\sum_{K=1,2}\ell_{FM}(\tilde{G}, D_K)\right], \tag{28}$$

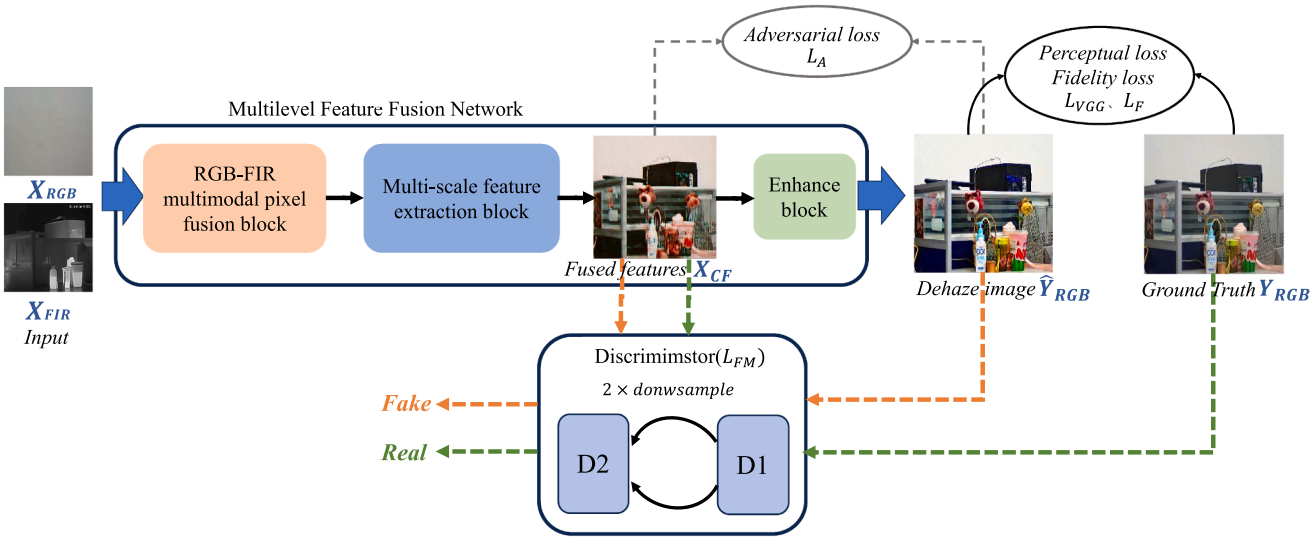


Fig. 6. RGB-FIR multimodal collaborative optimization learning process.

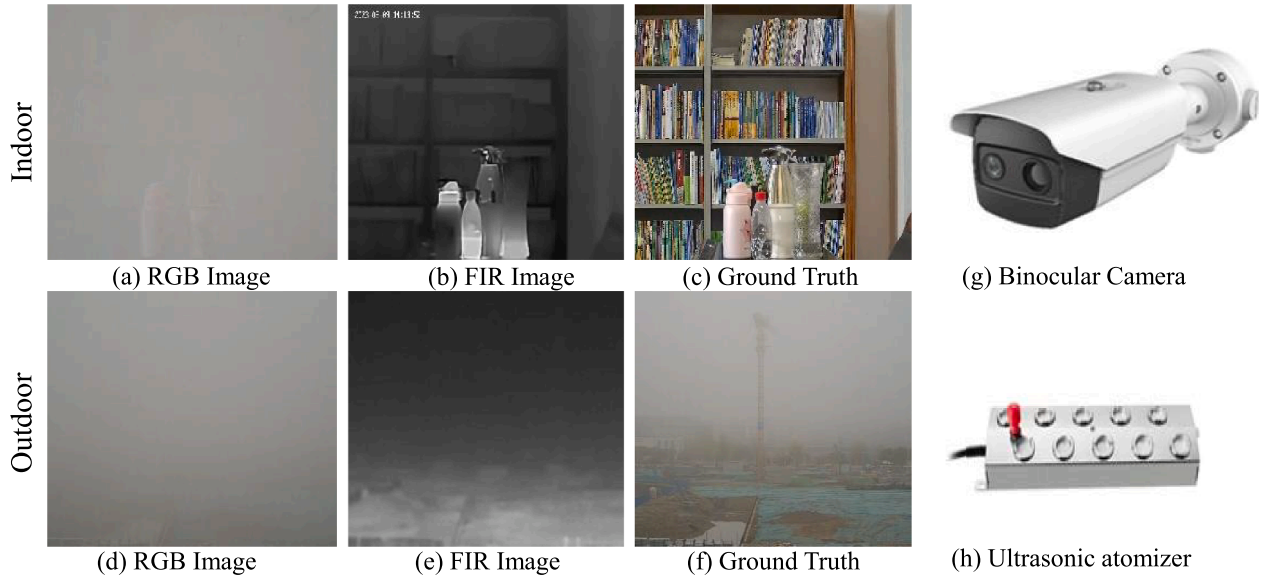


Fig. 7. An example of multimodal experimental data and the used acquisition devices.

$$\ell_{FM}(\tilde{G}, D_k) = E_{X_{RGB},Y_{FIR},Y_{RGB}} \sum_{i=1}^{T} \frac{1}{N_i} [\| D_k^{(i)}(Y_{RGB}) - D_k^{(i)}(\tilde{G}(X_{RGB}, X_{FIR}))\|_1], \tag{29}$$

where $\ell_{FM}(\tilde{G}, D_K)$ is the feature matching loss computed by the Kth discriminator D_k. T is the total number of layers for feature extraction. N_i is the number of elements in each layer. $D_k^{(i)}(\cdot)$ is the operator for feature extraction in the ith layer. $E_{X_{RGB},Y_{FIR},Y_{RGB}}$ is the expectation of the corresponding sample.

The perceptual loss L_{VGG} assesses the alignment between the defogged image and the ground truth, considering both perceptual and semantic aspects, to ensure that the defogged image closely matches the visual perception of human observers. Its precise definition is outlined below:

$$L_{VGG19}^{\phi,i}(\hat{Y}, Y_{RGB}) = \frac{1}{C_i H_i W_i} \|\phi_i(\hat{Y}_{RGB}) - \phi_i(Y_{RGB})\|_2^2, \tag{30}$$

where $\hat{Y}_{RGB}$ is the final result generated by the multimodal dehazing network. H_i and W_i are the height and width of the ith feature map. C_i is the number of channels. $C_i H_i W_i$ is the size of the ith layer feature map. $\phi_i(\hat{Y}_{RGB})$ is the feature representation of the ith layer in the VGG network. The fidelity loss L_F utilizes Euclidean distance to measure the difference between the ground-truth Y_{RGB} and the dehazed image $\hat{Y}_{RGB}$, which is defined as follows:

$$L_F = \|Y_{RGB} - \hat{Y}_{RGB}\|_2. \tag{31}$$

4. Experiments

4.1. Experimental settings

4.1.1. Datasets

Data acquisition. Due to the lack of publicly available RGB-FIR multimodal fog image datasets, we utilized a Hikvision binocular camera (DS-2TD2636T) capable of collecting RGB images with a resolution of 1920 × 1080 pixels and FIR images with a resolution of 384 × 384 pixels, respectively, where the infrared band covers 8 μm to 14 μm, as shown in Fig. 7(g). Simultaneously, we used an ultrasonic nebulizer to create

artificial fog, which is depicted in Fig. 7(h). Between September 2022 and December 2023, we collected a large amount of RGB-FIR multimodal fog data in various weather conditions with different fog concentrations. To ensure data diversity, we varied the shooting angle and distance in each of the 13 indoor and outdoor scenes where we deployed the binocular cameras. When collecting outdoor data, we chose locations prone to fog formation, such as factories, high-rise areas, and coastal regions. For indoor image collection, an ultrasonic atomizer was placed in front of the binocular camera, immersed in water until it produced fog, and then we adjusted the camera-atomizer distance to control fog density while capturing images with varying fog concentrations. We have successfully collected 11,736 pairs of foggy and corresponding clean images.

Image registration. To ensure consistency in viewing angle and resolution, we employ a straightforward image registration strategy. Initially, the high-resolution RGB image is downsampled. We utilize the Harris corner detection algorithm [26] to extract feature point information from the image. Subsequently, the Scale-Invariant Feature Transform (SIFT) [27] feature descriptor is employed to characterize these feature points and establish correspondences between the RGB and FIR images. Euclidean distance is then used to identify potential matches between the feature points of the two images. Following this, we apply the M-estimator Sample Consensus (MSAC) [28] algorithm to reject false matches and obtain a more precise transformation matrix, thereby enabling accurate registration of the RGB and FIR images. Finally, to facilitate model training, we adjust the resolution of the registered images using the bicubic algorithm in the OpenCV library, ensuring that the resolution and field of view of both images are consistent.

Dataset split. To ensure the validity of the model, the data for each scene is divided into a training set, a validation set, and a test set according to a ratio of 7:2:1. Both the training and validation sets comprise RGB-FIR image data pairs and corresponding truth images, whereas the test set contains only RGB-FIR image data pairs. In the comparative experiments, all models for comparison are retrained using the dataset described in this paper, with the image resolution set to 256×256.

4.1.2. Evaluation indicators

In evaluating the performance of dehazing algorithms, we utilize various image quality metrics and algorithm complexity indicators for a comprehensive analysis. Specifically, in reference to DehazeNet [9] and EPDN [13], we adopt Mean Squared Error (MSE), Structural Similarity Index Measure (SSIM) and Peak signal-to-noise ratio (PSNR)to evaluate the dehazing effectiveness. Color saturation (SAT) [29] evaluates the vividness of colors in the image. Additionally, the Frechet Inception Distance (FID) [30] and Learned Perceptual Image Patch Similarity (LPIPS) [31] are utilized to measure the quality of dehazed images, based on their proficiency in evaluating the quality of generated images. These metrics evaluate the performance of dehazing algorithms from different perspectives in a comprehensive manner. Additionally, we consider the number of model parameters (Params, M) and the computational cost (FLOPs, G) as indicators of algorithm complexity and real-time performance to assess the efficiency and practicality of each model in dehazing image processing.

4.1.3. Implementation details

This work employs Python 3.9 and Pytorch 1.10 to construct the framework for an RGB-FIR multimodal fusion-transfer network. The server environment is as follows: Windows 11, CPU (Intel Core i7-11800H), GPU (NVIDIA GeForce RTX3060), RAM (32 GB), CUDA 11.3. The model is configured with a learning rate set to 0.0002, a patch size of 70×70, a batch size of 4, and a β_1 value of 0.6, all using the Adam optimizer.

4.2. Ablation studies

4.2.1. Setting of parameters λ_1, λ_2, and λ_3

To determine the values of parameters λ_1, λ_2, and λ_3, we conduct an ablation study to investigate the individual influence of the constraints in Eq. (25). Inspired by the pioneering research of GANs in image style transfer [13,32], we set the balancing weights for the adversarial loss and feature matching loss to 1 and 100, respectively (i.e., the value of parameter λ_1 is 100). To enhance the precision of our defogging model in restoring the intricate image structure and content details, we have undertaken an exhaustive evaluation of the optimal settings for the weight parameters, λ_2 and λ_3, which govern the perceptual loss and

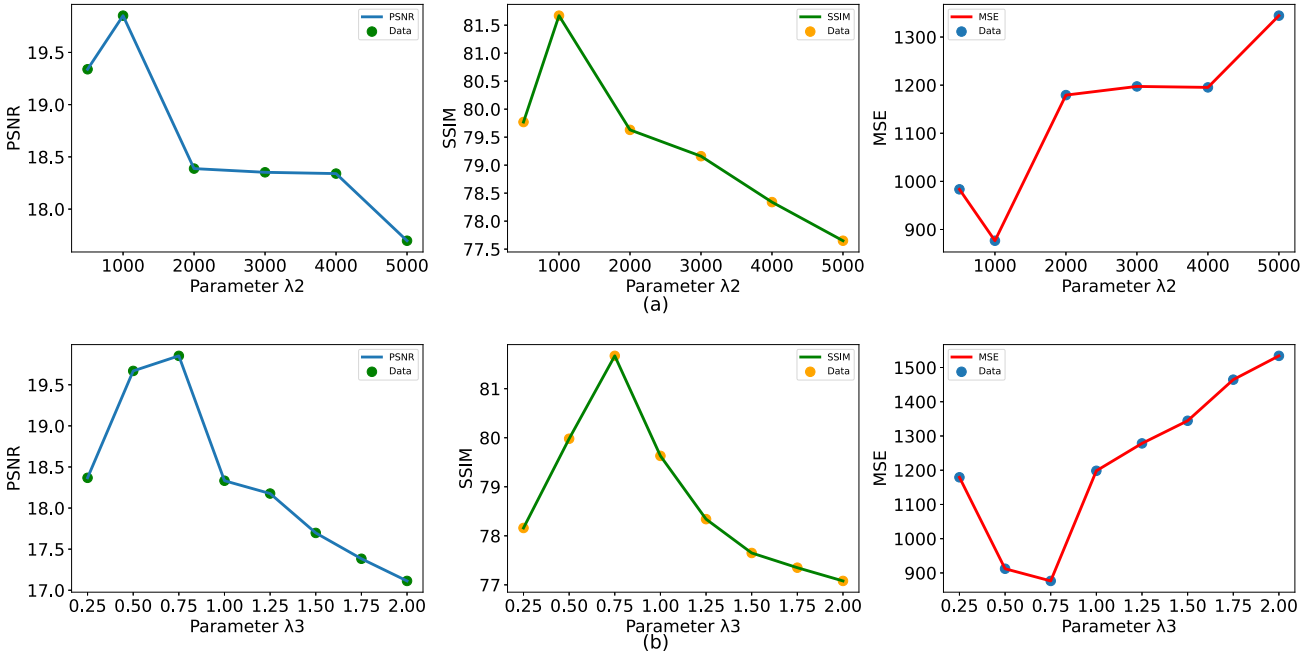


Fig. 8. The ablation studies of parameters λ_2 and λ_3.

Table 1
The impact of single-modal vs multimodal. The best results are shown in bold.

Method Modal/Training Method	Outdoor			Indoor		
	SSIM↑	PSNR↑	MSE↓	SSIM↑	PSNR↑	MSE↓
RGB Single-modal/EPDN [13]	0.7627	17.3878	1464.5068	0.7524	17.0652	1567.0367
FIR Single-modal/EPDN [13]	0.7989	17.6923	1344.4389	0.7611	17.3659	1480.2883
RGB-FIR Multimodal/P2P [32]	0.8012	17.3762	1475.2883	0.7934	18.5233	1132.2476
RGB-FIR Multimodal/Ours	**0.8381**	**18.6681**	**1115.2476**	**0.8202**	**20.3275**	**722.2457**

Table 2
The impact of the different attention modules. The best results are shown in bold.

Method	Outdoor			Indoor		
	SSIM↑	PSNR↑	MSE↓	SSIM↑	PSNR↑	MSE↓
No attention + Multimodal	0.8273	18.0500	1295.7413	0.8028	19.2734	983.4577
CBAM + Multimodal	0.8279	17.8312	1318.8726	0.8039	18.3262	1200.8816
ECA + Multimodal	0.8316	17.8711	1304.2295	0.8074	18.3400	1197.3149
SE-Net + Multimodal	0.8319	18.1765	1278.6546	0.8124	19.6517	912.6742
CMCA + Multimodal(Ours)	**0.8381**	**18.6681**	**1115.2476**	**0.8202**	**20.3275**	**722.2457**

fidelity loss, respectively. Inspired by the GAN-based image enhancement research [33], we have carefully calibrated the weight parameter λ_2 associated with the perceptual loss, setting its order of magnitude to 10^3. To determine the optimal value, we conducted a series of ablation experiments, spanning the range from 500 to 5000, with increments of 1000. The resulting analysis, depicted in Fig. 8(a), reveals a clear pattern: when λ_2 is set to 1000, our model achieves optimal performance across the key metrics of PSNR, SSIM, and MSE. This finding is attributed to the balance achieved between two opposing factors: an overly large λ_2 diminishes the inference capability of our three-branch generator, while an insufficiently weighted λ_2 leads to a noticeable blurriness in the enhanced images. Drawing upon the findings of the study [13], we conducted a rigorous ablation study to determine the optimal setting for the weight parameter λ_3 in the fidelity loss function. With 1 as the baseline, we varied λ_3 within the range of [0.25, 2.5], using increments of 0.25. The results are presented in Fig. 8(b), which clearly indicates that when λ_3 is set to 0.75, the defogging model achieves peak performance.

4.2.2. Single-modal dehazing VS multimodal dehazing

Here, we conduct an ablation study to validate the effectiveness of our proposed RGB-FIR multimodal fusion-transfer dehazing network. Specifically, we compare four dehazing networks: an RGB-based single-modal dehazing network, an FIR-based single-modal dehazing network, an RGB-FIR multimodal dehazing network based on Pix2Pix training method, and our RGB-FIR multimodal dehazing network. The single-modal networks utilize the multi-scale feature-level fusion module and the image transfer module, while the multimodal network employs the pixel-level fusion module, the multi-scale feature-level fusion module, and the image transfer module. Fig. 9 shows a subjective visual comparison of the four networks. From this, we can observe that the images enhanced by the multimodal dehazing network exhibit clearer and richer details and texture information. This result directly demonstrates that the proposed RGB-FIR multimodal dehazing network effectively fuses RGB and FIR image features, achieving modal complementarity, which significantly enhances the dehazing effect. In Table 1, we showcase the MSE, SSIM, and PSNR test results of the four dehazing networks evaluated on a self-built dehazing dataset. The results reveal a clear superiority of our proposed RGB-FIR multimodal dehazing network, which surpasses the RGB, FIR single-modal dehazing model and the Pix2Pix-based multimodal dehazing model across MSE, SSIM, and PSNR metrics.

4.2.3. Ablation studies of CMCA

To validate the improvement effect of the proposed CMCA module on the multimodal dehazing model, we compare the RGB-FIR multimodal dehazing model without an attention mechanism, as well as RGB-FIR multimodal dehazing models based on CBAM, ECA, SE-Net, and our proposed CMCA. The experimental results are presented in Table 2. On outdoor scene data, CMCA has achieved a 1.08 % improvement in SSIM, an enhancement of 0.6181 dB in PSNR, and a reduction of 180.4937 in MSE. However, its performance is even more significant on indoor scene data, delivering a 1.74 % increase in SSIM, a 1.0541 dB enhancement in PSNR, and a substantial reduction of 261.2120 in MSE. The experimental results indicate that the proposed CMCA mechanism is capable of capturing spatial importance weights from FIR pixel-level features that possess “fog penetration” capabilities, thereby precisely optimizing the low-level pixel features at corresponding spatial positions in RGB images to achieve cross-modal complementary advantages. On the contrary, the dehazing models based on the CBAM and ECA attention mechanisms have seen a decrease in the PSNR and MSE metrics, which suggests that the feature-level channel and spatial attention models, such as CBAM and ECA, are not suitable for optimizing pixel-level features at the low-level. Fig. 10 presents the visual examples of RGB-FIR fused features (i.e. X''_{fusion}) for two samples, modulated by different attention models. Notably, the contours and shapes of some objects within the X''_{fusion} maps resulting from CMCA can be more clearly and effectively displayed compared to other competing attention mechanisms.

In the CMCA module, the integration of various scale pooling factors significantly influences the efficacy of the RGB-FIR multimodal defogging model. To ascertain the ideal pooling factor combination for maximizing defogging performance, we conduct a thorough quantitative evaluation, whose outcomes are outlined in Table 3. Upon comparing the MSE, SSIM, and PSNR metrics across different scale pooling factor combinations, we have drawn the following conclusions:

(1) Collectively, the models incorporating multi-scale pooling combinations outperform their single-scale counterparts. This observation

Table 3
The impact of the CMCA module under different pooling factors. The best results are shown in **bold**.

Method	Combination	SSIM↑	PSNR↑	MSE↓
Single-scale	(1,3)	0.8331	18.3944	1174.3149
	(1,5)	0.8327	18.3523	1189.5859
	(1,7)	0.8323	18.3237	1200.9612
Multi-scale	(1,3) + (1,5)	0.8355	18.4938	1161.4055
	(1,5) + (1,7)	0.8335	18.3275	1200.6525
	(1,3) + (1,7)	**0.8381**	**18.6681**	**1115.2476**
	(1,3) + (1,5) + (1,7)	0.8339	18.3433	1197.2165

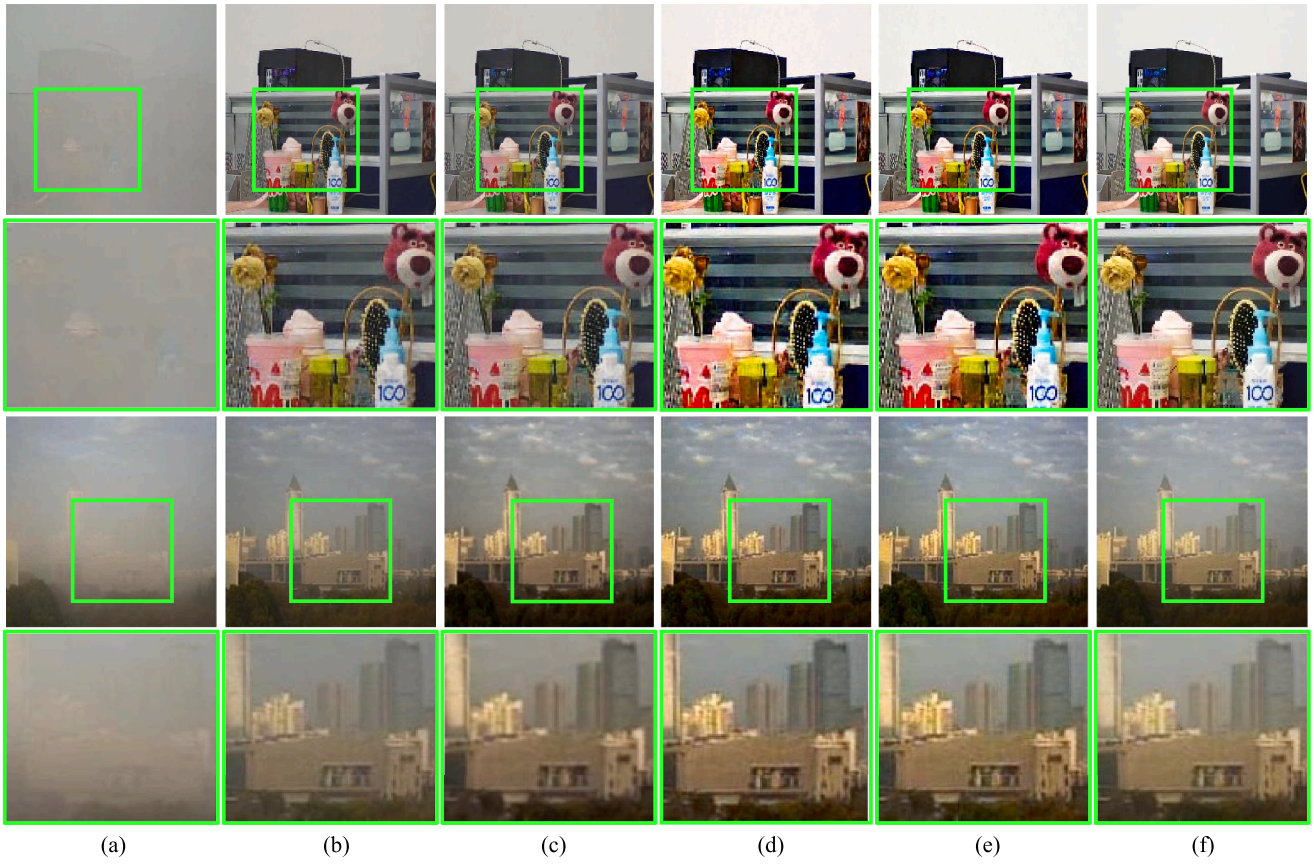


Fig. 9. Example of the results for single-modal dehazing vs multimodal dehazing. (a) are hazy images. (b) are the results of RGB single-modal dehazing. (c) are the results of FIR single-modal dehazing. (d) are the results of multimodal dehazing (P2P-based training method). (e) are the results of multimodal dehazing (Our). (f) are Ground-truth images.

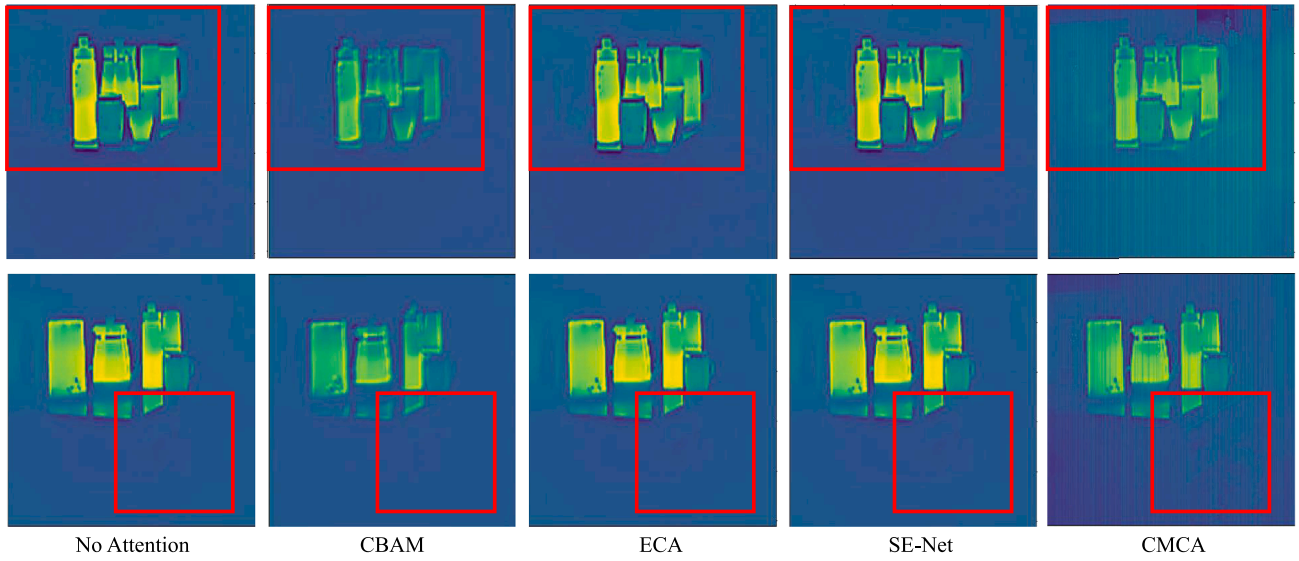


Fig. 10. Comparison of adjusted feature maps of each attention mechanism.

indicates that the integration of pooling operations across diverse scales allows for a more nuanced capture of multi-scale features in images, ultimately enhancing the defogging outcome.

(2) Notably, when the pooling factor combination is set to (1,3) and (1,7), the multimodal defogging network model attains superior defogging capabilities. In comparison to a model utilizing only a single scale of (1,7), this combined configuration achieves a noteworthy enhancement of 0.0057 in SSIM, a significant increase of 0.3444 dB in PSNR, and a substantial reduction of 85.7136 in MSE.

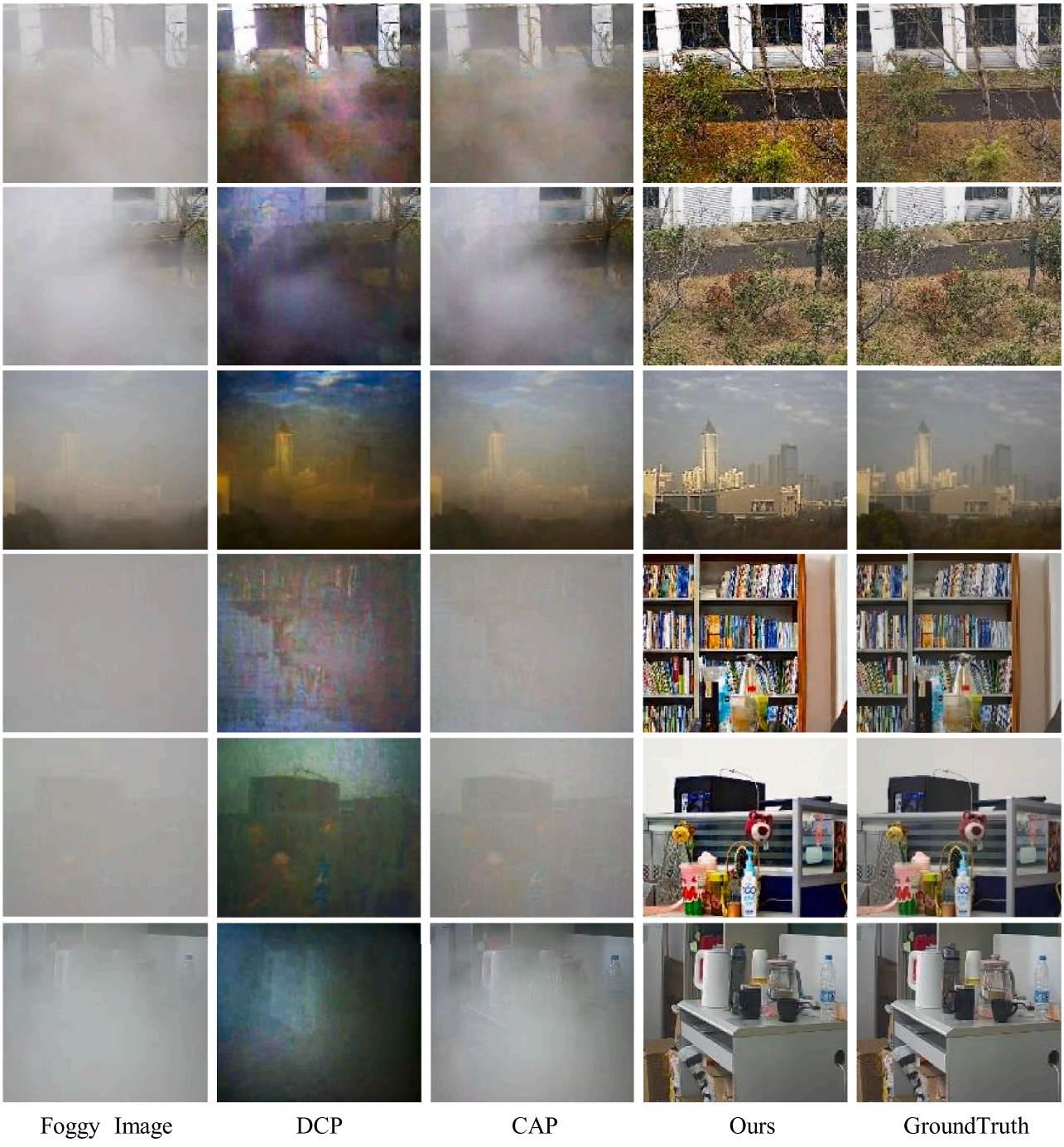


Fig. 11. Examples of prior knowledge-based dehazing algorithms.

4.3. Comparisons with state-of-the-arts

To validate the effectiveness of the proposed RGB-FIR multimodal dehazing algorithm, we have selected thirteen popular dehazing algorithms as competing algorithms, including two prior-based dehazing algorithms, DCP [4] and CAP [7], as well as eleven deep learning-based dehazing algorithms, DehazeNet [9], Pix2Pix [32], EPDN [13], DAD [34], MB-TF [35], D4Net [36], Dehaze-former (DHF) [37], DeHamer (DH) [11], C2PNet [38], UCL-Dehaze(UCL) [39] and VIFNet(VIF) [40]. Table 4 provides a detailed comparison of eight evaluation metrics for fourteen dehazing algorithms tested on our self-built dataset. The results show that Pix2Pix [32], EPDN [13], and the proposed method stand out significantly in terms of MSE, SSIM, PSNR, FID, and LPIPS. This strongly suggests that dehazing methods based on image style transfer can effectively improve dehazing performance under the collaborative optimization of multiple losses, such as adversarial generation loss, VGG perceptual loss, and fidelity loss. Furthermore, our method outperforms both Pix2Pix and EPDN in evaluation metrics including MSE, SSIM, PSNR, SAT, FID, and LPIPS. This advantage mainly comes from two key factors:

- The RGB-FIR multimodal dehazing network is based on RGB and FIR dual-input images, providing richer feature information for the dehazing process.
- The designed multimodal fusion-transfer module realizes multi-level modal complementarity at pixel and feature levels, and effectively

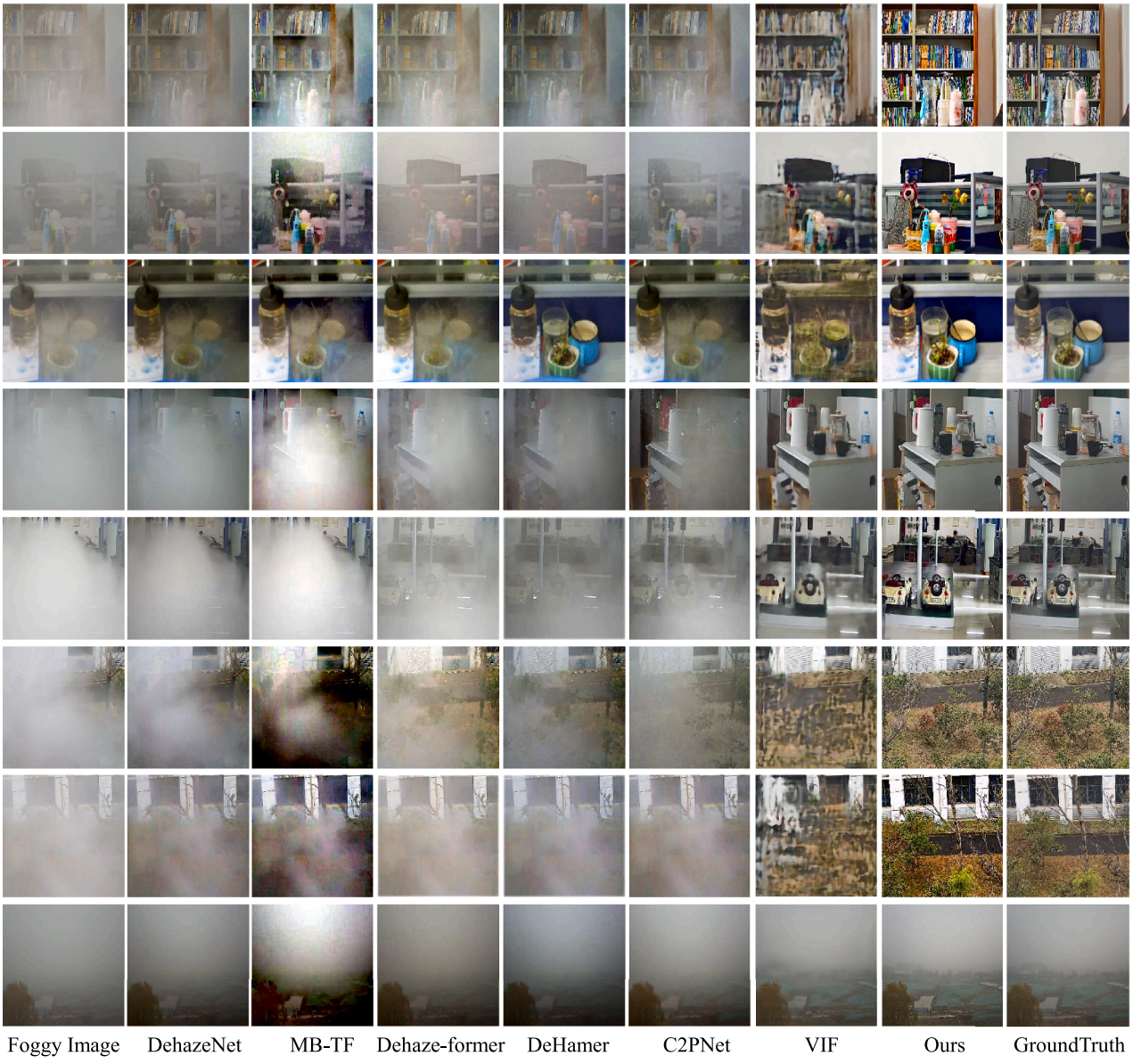


Fig. 12. Examples of some deep learning-based dehazing algorithms.

adjusts the network weights between the multi-level fusion module and the multi-scale image transfer module through a collaborative optimization strategy.

The bottom two rows of Table 4 present a comparison of different dehazing algorithms in terms of network parameters and computational complexity. Among various deep learning algorithms, DehazeNet [9] excels in terms of computation and total number of parameters, yet lags significantly in defogging performance. The three algorithms with the best defogging effects, Pix2Pix, EPDN, and the proposed method, all have moderate levels of parameters and computational complexity among all competing algorithms, which indicates that our algorithm has achieved a considerable balance in terms of parameter count, computational efficiency, and defogging performance.

Figs. 11–13 visually compare the subjective dehazing performance of eleven dehazing algorithms, including those based on prior knowledge, deep learning, and image style transfer, with the proposed dehazing method in real foggy weather scenarios. As shown in Fig. 11, the DCP [4] and CAP [7] algorithms still demonstrate some effectiveness in lightly hazy scenes. However, when faced with dense fog, their performance becomes inadequate, and the processed images remain visually similar to the original hazy images. In contrast, the proposed method is able to effectively restore the texture details and color information of hazy images.

As highlighted in Fig. 12, the comparison of subjective dehazing performance between our proposed method and various deep learning-based algorithms underscores the superiority of our approach, particularly in dense foggy scenes. While the deep learning algorithms, such as MB-TF [35], and DeHamer [11], demonstrate some effectiveness in light fog conditions and are able to restore certain details obscured by haze, their performance significantly diminishes in dense fog. However, when compared to VIFNet [40], our method performs better in dense foggy scenarios, effectively restoring color and detail information lost in hazy images. As shown in Fig. 12, our method produces dehazed images with higher color saturation, enhanced contrast, and clearer detail features, outperforming other deep learning-based methods. This demonstrates the robustness and effectiveness of the proposed method in handling challenging dehazing tasks, especially in dense foggy environments.

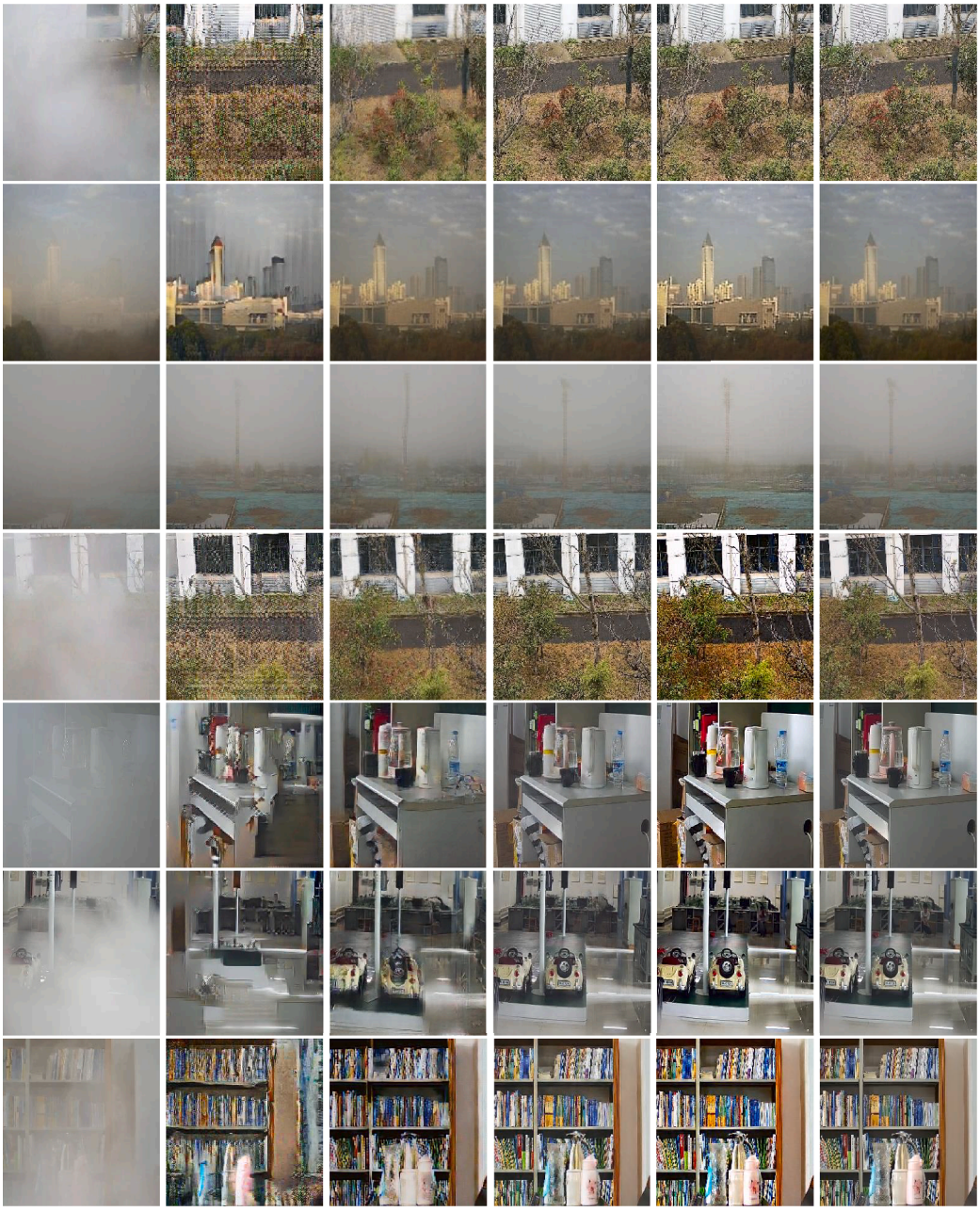


Fig. 13. Examples of image style transfer based-dehazing algorithms.

As depicted in Fig. 13, the dehazing algorithms UCL [39], Pix2Pix [32], and EPDN [13], while exhibiting some enhancement in image contrast and color saturation, fall short in comparison to the remarkable superiority of our proposed algorithm. Our method not only successfully removes the haze from the image but also significantly improves the contrast and color saturation, enabling distant scenery to clearly reveal its appearance and colors. This superiority stems from the effective utilization of the complementary relationship between RGB and FIR features in foggy conditions by our proposed algorithm, thereby enhancing the features of the scenery covered by fog.

4.4. Model generalizability evaluation

To verify the robustness of the proposed model in unknown haze scenarios, model generalization experiments are designed. Given the current absence of publicly available real-world RGB-FIR dehazing datasets, we have validated the generalization capability of our proposed model using our self-constructed dataset containing both indoor and outdoor scenarios. The experimental setup was designed as follows: (1) Cross-domain testing I: Evaluating the model trained on outdoor data for indoor scene dehazing (i.e., Outdoor-to-Indoor Transfer). (2) Cross-domain testing II: Evaluating the model trained on indoor data for outdoor scene dehazing (i.e., Indoor-to-Outdoor Transfer). Furthermore, we have conducted comprehensive performance comparisons with several SOTA dehazing models and provided quantitative analysis results. Table 5 shows the generalizability comparison results. Obviously, the dehazing performance of all the models is degraded in unknown scenarios. However, our model has a stronger generalization ability compared to other competing models. This enhancement primarily comes from two key factors: the effective utilization of FIR images and the integration

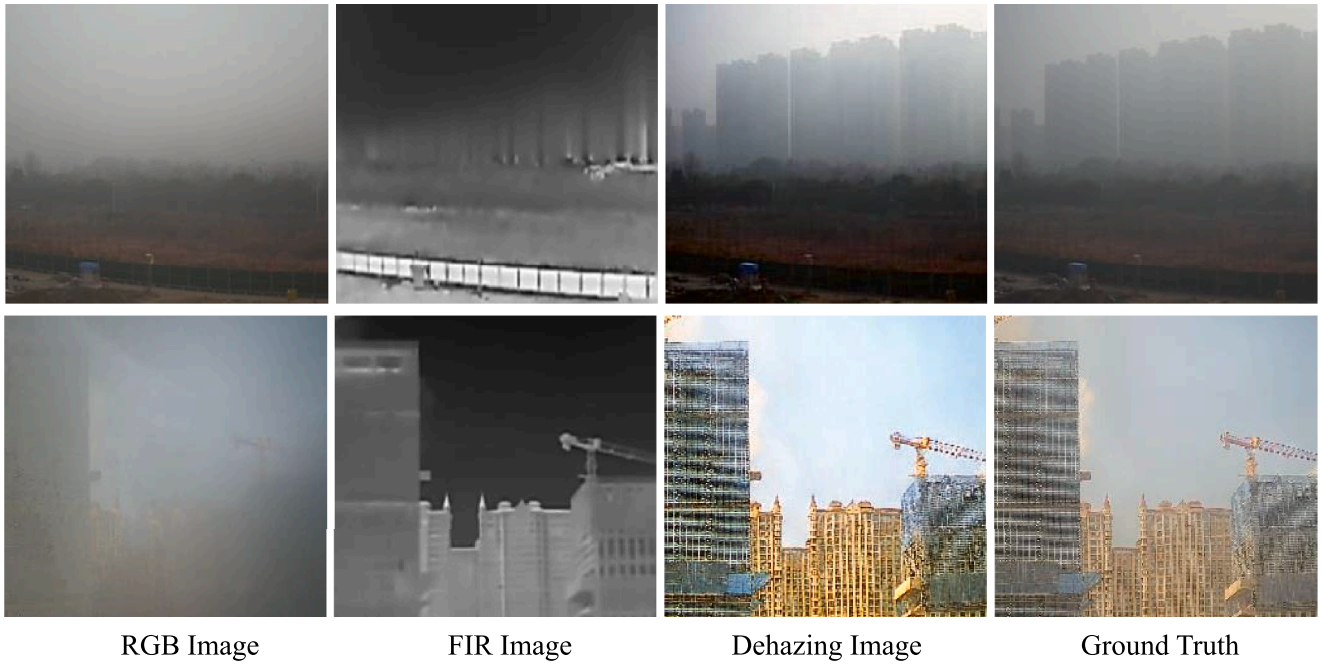


Fig. 14. Examples of some unsatisfactory enhancement results from the proposed algorithm.

Table 4
Objective visual quality evaluations of our method and other SOTA methods using PSNR, SSIM, MSE, SAT, FID, LPIPS, Params, FLOPs. The best results are shown in **bold**.

Index	DCP [4]	CAP [7]	Dehaze [9]	PixPix [32]	MB-TF [35]	D4Net [36]	EPDN [13]	DAD [34]	DHF [37]	DH [11]	C2PNet [38]	UCL [39]	VIF [40]	Ours
Indoor														
SSIM↑	0.48	0.53	0.57	0.72	0.57	0.52	0.76	0.66	0.64	0.61	0.65	0.66	0.55	**0.82**
PSNR↑	13.71	13.06	14.41	17.64	15.93	13.32	18.42	14.87	14.83	13.46	13.45	17.19	18.37	**20.32**
MSE↓	2995.1	3384.4	2535.2	1362.5	1658.3	3141.1	1142.9	2117.5	2143.8	2926.9	2936.4	1554.5	1182.5	**722.2**
SAT↑	0.30	0.12	0.13	0.22	0.19	0.13	0.20	0.17	0.14	0.14	0.17	0.24	0.22	**0.29**
FID↓	175.1	180.1	67.7	35.2	61.4	80.7	28.9	75.9	73.4	71.2	80.2	99.7	121.1	**27.6**
LPIPS↓	0.82	0.86	0.58	0.24	0.42	0.70	0.20	0.52	0.60	0.61	0.56	0.49	0.48	**0.16**
Outdoor														
SSIM↑	0.37	0.49	0.49	0.75	0.44	0.43	0.81	0.48	0.62	0.59	0.63	0.71	0.53	**0.83**
PSNR↑	13.18	13.28	14.81	17.24	14.58	13.66	18.14	15.32	14.84	14.01	14.32	17.57	17.73	**18.66**
MSE↓	3230.5	3198.1	2152.8	1505.5	2358.1	3043.7	1285.6	1923.8	2132.1	2503.9	2778.7	1399.5	1332.4	**1115.2**
SAT↑	0.21	0.14	0.13	0.21	0.17	0.12	0.22	0.21	0.16	0.14	0.15	0.23	0.20	**0.26**
FID↓	183.1	189.4	117.2	60.1	125.3	135.8	30.7	121.2	96.2	102.5	89.6	100.8	132.0	**28.4**
LPIPS↓	0.71	0.76	0.80	0.34	0.69	0.79	0.17	0.72	0.65	0.63	0.69	0.47	0.76	**0.17**
Params↓	–	–	**0.01**	22.67	7.43	10.7	17.38	15.64	1.28	132.45	7.17	19.45	9.78	22.23
FLOPs↓	–	–	**0.6**	10.54	2.24	2.25	4.82	13.97	13.13	60.15	2.18	6.75	2.42	7.82

Table 5
Generalizability performance comparison of dehazing methods. The best results are shown in **bold.**

Method	Outdoor-to-indoor transfer			Indoor-to-outdoor transfer		
	SSIM↑	PSNR↑	MSE↓	SSIM↑	PSNR↑	MSE↓
VIF [40]	0.52	16.38	1622.7	0.51	15.75	1735.7
Pix2Pix [32]	0.63	14.76	2197.9	0.66	14.47	2472.8
EPDN [13]	0.66	15.44	1869.4	0.71	15.21	2013.7
UCL [39]	0.61	14.35	2697.6	0.62	14.79	2171.2
Ours	**0.71**	**16.89**	**1581.5**	**0.73**	**15.96**	**1656.1**

of multimodal pixel-level and multi-scale feature-level fusion modules. These components strengthen cross-modal complementarity, allowing the model to extract critical structural and semantic information even in dense fog. As a result, the model achieves better generalization and superior defogging performance in unseen scenarios.

4.5. Discussion

Fig. 14 illustrates several instances where the proposed algorithm exhibits unsatisfactory enhancement performance. We analyze the reasons as follows: on the one hand, during the data collection stage, the weather environment may not have been completely clear when collecting certain fog-free image data, resulting in the original image containing a small amount of fog. Therefore, the fog-free image corresponding to the foggy image is not completely pure, and during the learning process, the model cannot really learn the features of the data in the fog-free state, leading to ineffective final output of the defogged image. On the other hand, the camera used in this paper's method is based on the principle of thermal imaging, which expresses the features through the heat difference of the object. Consequently, the imaging difference caused by the temperature difference of the object is also a factor affecting the model's de-fogging effect. From the FIR map, it can be seen that when the temperature difference of some people or objects in the FIR image is not large, the feature information extracted from it may not be

sufficient, and the fusion features output in the fusion stage cannot provide sufficiently rich FIR information, resulting in poor de-fogging effect in some places where the temperature difference is not obvious.

In terms of applicable scenarios, the multimodal dehazing network based on the adversarial generation strategy offers the dual advantages of information fusion and adversarial generation. This makes it highly valuable for applications in various fields such as traffic monitoring, unmanned aerial vehicle (UAV) image processing, remote sensing and monitoring, and security monitoring. Specifically, in fields like traffic monitoring, UAV image processing, and remote sensing and monitoring, where visual task performance can be significantly hindered by haze, this network's effectiveness is particularly notable.

5. Conclusion

In addressing the inherent challenges of existing methods for processing RGB images in dense fog scenarios, we introduce a novel RGB-FIR multimodal fusion-transfer network. Our method integrates RGB and FIR image information, thereby substantially enriching the input data and overcoming the limitations of conventional RGB-only methods. Specifically, the novel fusion-transfer network, along with its collaborative optimization learning strategy, successfully establishes a robust nonlinear mapping relationship between RGB-FIR multimodal fused features and ground-truth images. To validate the effectiveness of the proposed method, we have compiled a comprehensive dataset consisting of 11,736 pairs of foggy and foggy-free RGB-FIR multimodal images. Experimental results demonstrate that our multimodal dehazing method outperforms current popular methods in various evaluation metrics, including MSE, SSIM, PSNR, and SAT, particularly when dealing with images captured in dense fog environments. In future research, we will delve deeper into methods for multimodal data processing and feature fusion, with a focus on enhancing model robustness and compressing the parameter size.

CRediT authorship contribution statement

Ruolin Du: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation; **Han Wang:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization; **Wenjie Liu:** Investigation, Funding acquisition, Formal analysis, Data curation; **Guangcheng Wang:** Writing – review & editing, Visualization, Validation, Supervision, Formal analysis; **Kui Jiang:** Writing – review & editing, Visualization, Supervision, Formal analysis; **Hanseok Ko:** Supervision, Resources, Project administration, Formal analysis.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial-interestsor personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61872425, 62401296, 62301287, and in part by the Nantong Natural Science Foundation under Grant JC2023005.

References

[1] Y. Liu, X. Hou, Local multi-scale feature aggregation network for real-time image dehazing, Pattern Recognit. 141 (2023). https://doi.org/10.1016/j.patcog.2023.109599
[2] N. Jiang, K. Hu, T. Zhang, W. Chen, Y. Xu, T. Zhao, Deep hybrid model for single image dehazing and detail refinement, Pattern Recognit. 136 (2023). https://doi.org/10.1016/j.patcog.2022.109227
[3] F. Yuan, Y. Zhou, X. Xia, X. Qian, J. Huang, A confidence prior for image dehazing, Pattern Recognit. 119 (2021). https://doi.org/10.1016/j.patcog.2021.108076
[4] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1956–1963. https://doi.org/10.1109/CVPR.2009.5206515
[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144. https://doi.org/10.1145/3422622
[6] T. Senior, D. Ahlgren, Rayleigh scattering, IEEE Trans. Antennas Propag. 21 (1) (1973) 134. https://doi.org/10.1109/TAP.1973.1140429
[7] Q.S. Zhu, J.M. Mai, L. Shao, A fast single image haze removal algorithm using color attenuation prior, IEEE Trans. Image Process. 24 (11) (2015) 3522–3533. Times Cited in Web of Science Core Collection: 1286 Total Times Cited: 1476. https://doi.org/10.1109/TIP.2015.2446191
[8] S.K. Yadav, K. Sarawadekar, Robust multi-scale weighting-based edge-smoothing filter for single image dehazing, Pattern Recognit. 149 (2024). https://doi.org/10.1016/j.patcog.2023.110137
[9] B.L. Cai, X.M. Xu, K. Jia, C.M. Qing, D.C. Tao, DehazeNet: an end-to-end system for single image haze removal, IEEE Trans. Image Process. 25 (11) (2016) 5187–5198. Times Cited in Web of Science Core Collection: 1641 Total Times Cited: 1875. https://doi.org/10.1109/TIP.2016.2598681
[10] H. Zhang, V.M. Patel, Densely connected pyramid dehazing network, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3194–3203. https://doi.org/10.1109/CVPR.2018.00337
[11] C. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, C. Li, Image dehazing transformer with transmission-aware 3D position embedding, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5802–5810. https://doi.org/10.1109/CVPR52688.2022.00572
[12] Y. Wang, J. Xiong, X. Yan, M. Wei, USCFormer: unified transformer with semantically contrastive learning for image dehazing, IEEE Trans. Intell. Transp. Syst. 24 (10) (2023) 11321–11333. https://doi.org/10.1109/TITS.2023.3277709
[13] Y. Qu, Y. Chen, J. Huang, Y. Xie, Enhanced Pix2pix dehazing network, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8152–8160. https://doi.org/10.1109/CVPR.2019.00835
[14] D. Engin, A. Genc, H.K. Ekenel, Cycle-dehaze: enhanced CycleGAN for single image dehazing, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 938–9388. https://doi.org/10.1109/CVPRW.2018.00127
[15] Y. Wang, X. Yan, D. Guan, M. Wei, Y. Chen, X.-P. Zhang, J. Li, Cycle-SNSPGAN: towards real-world image dehazing via cycle spectral normalized soft likelihood estimation patch GAN, IEEE Trans. Intell. Transp. Syst. 23 (11) (2022) 20368–20382. https://doi.org/10.1109/TITS.2022.3170328
[16] T. Shibata, M. Tanaka, M. Okutomi, Visible and near-infrared image fusion based on visually salient area selection, Proc. SPIE Int. Soc. Opt. Eng. 9404 (2015). https://doi.org/10.1117/12.2077050
[17] R. Kumar, B.K. Kaushik, R. Balasubramanian, Multispectral transmission map fusion method and architecture for image dehazing, IEEE Trans. Very Large Scale Integr. VLSI Syst. 27 (11) (2019) 2693–2697. https://doi.org/10.1109/TVLSI.2019.2932033
[18] J. Liu, G. Wu, Z. Liu, D. Wang, Z. Jiang, L. Ma, W. Zhong, X. Fan, R. Liu, Infrared and visible image fusion: from data compatibility to task adaption, IEEE Trans. Pattern Anal. Mach. Intell. 47 (4) (2025) 2349–2369. https://doi.org/10.1109/TPAMI.2024.3521416
[19] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, L. Van Gool, Equivariant multi-modality image fusion, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 25912–25921. https://doi.org/10.1109/CVPR52733.2024.02448
[20] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, L. Van Gool, CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5906–5916. https://doi.org/10.1109/CVPR52729.2023.00572
[21] J. Liu, R. Lin, G. Wu, R. Liu, Z. Luo, X. Fan, CoCoNet: coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion, Int. J. Comput. Vis. 132 (5) (2024) 1748–1775. https://doi.org/10.1007/s11263-023-01952-1
[22] J. Liu, G. Wu, J. Luan, Z. Jiang, R. Liu, X. Fan, HoLoCo: holistic and local contrastive learning network for multi-exposure image fusion, Inf. Fusion 95 (2023) 237–249. https://doi.org/10.1016/j.inffus.2023.02.027
[23] W. Di, L. Jinyuan, L. Risheng, F. Xin, An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection, Inf. Fusion 98 (2023) 101828. https://doi.org/10.1016/j.inffus.2023.101828
[24] R. Liu, Z. Liu, J. Liu, X. Fan, Z. Luo, A task-guided, implicitly-searched and meta-initialized deep model for image fusion, IEEE Trans. Pattern Anal. Mach. Intell. 46 (10) (2024) 6594–6609. https://doi.org/10.1109/TPAMI.2024.3382308
[25] D. Wang, J. Liu, L. Ma, R. Liu, X. Fan, Improving misaligned multi-modality image fusion with one-stage progressive dense registration, IEEE Trans. Circuits Syst. Video Technol. 34 (11) (2024) 10944–10958.

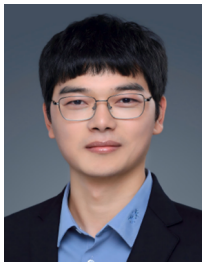
[26] C.G. Harris, M.J. Stephens, A combined corner and edge detector, in: Alvey Vision Conference, 1988.
[27] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, 2, 1999, pp. 1150–1157. https://doi.org/10.1109/ICCV.1999.790410
[28] P.H.S. Torr, A. Zisserman, MLESAC: a new robust estimator with application to estimating image geometry, Comput. Vis. Image Underst. 78 (1) (2000) 138–156. https://doi.org/10.1006/CVIU.1999.0832
[29] S.-L. Lee, C.-C. Tseng, A back lighting color image enhancement method using color saturation and image fusion, in: 2015 IEEE International Conference on Consumer Electronics - Taiwan, 2015, pp. 23–24. https://doi.org/10.1109/ICCE-TW.2015.7216870
[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, ArXiv pre-print (2017) 6629. https://doi.org/10.48550/arXiv.1706.08500
[31] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595. https://doi.org/10.1109/CVPR.2018.00068
[32] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5967–5976. https://doi.org/10.1109/CVPR.2017.632
[33] W. Han, Y. Xiao, Y. Yin, UM-GAN: Underground mine GAN for underground mine low-light image enhancement, IET Image Proc. (2024). https://doi.org/10.1049/ipr2.13092
[34] Y. Shao, L. Li, W. Ren, C. Gao, N. Sang, Domain adaptation for image dehazing, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2805–2814. https://doi.org/10.1109/CVPR42600.2020.00288
[35] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, Z. Jin, MB-TaylorFormer: multi-branch efficient transformer expanded by Taylor formula for image dehazing, in: International Conference on Computer Vision (ICCV), 2023. https://arxiv.org/abs/2308.14036
[36] G. Huang, Z. Liu, L.V.D. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243
[37] Y. Song, Z. He, H. Qian, X. Du, Vision transformers for single image dehazing, IEEE Trans. Image Process. 32 (2023) 1927–1941. https://doi.org/10.1109/TIP.2023.3256763
[38] Y. Zheng, J. Zhan, S. He, J. Dong, Y. Du, Curricular contrastive regularization for physics-aware single image dehazing, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5785–5794. https://doi.org/10.1109/CVPR52729.2023.00560
[39] Y. Wang, X. Yan, F.L. Wang, H. Xie, W. Yang, X.-P. Zhang, J. Qin, M. Wei, UCL-dehaze: toward real-world image dehazing via unsupervised contrastive learning, IEEE Trans. Image Process. 33 (2024) 1361–1374. https://doi.org/10.1109/TIP.2024.3362153
[40] M. Yu, T. Cui, H. Lu, Y. Yue, VIFNet: an end-to-end visible–infrared fusion network for image dehazing, Neurocomputing 599 (2024) 128105. https://doi.org/10.1016/j.neucom.2024.128105

Ruolin Du received the M.S. degree in Artificial Intelligence from the School of Transportation and Civil Engineering, Nantong University in 2025. Her research interest includes computer vision in intelligent transportation field, including multimodal image processing, feature fusion, transfer learning, and deep learning.

Han Wang received the M.S. degree from Northeastern University in 2007, and the Ph.D. degree in electrical engineering from Korea University in 2015. He iscurrently an Assistant Professor with the School of Transportation and Civil Engineering, Nantong University. His current research interests include multimodal imageprocessing, computer vision.

Wenjie Liu received the Ph.D. degree in systems innovation engineering from Tokushima University, Japan, in 2021. He is currently a Research Assistant at Nantong University, China. His research interests include image analysis, computer vision, and artificial intelligence.

Guangcheng Wang received the Ph.D. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2022. He is currently an Associate Professor with the School of Transportation and Civil Engineering, Nantong University, Nantong, China. His research interests include image processing and machine learning.



Kui Jiang (Member, IEEE) received the Ph.D. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2022. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His research interests include image/video processing and computer vision.

Hanseok Ko (Senior Member, IEEE) received the Ph.D. degree from the School of Electrical Engineering, Catholic University of America in 1992. He is currently a Professor with the Faculty of the Department of Electronics and Computer Engineering, Korea University. His research interests include multimodal sensor fusion for human–computer interaction.