DFG-DDM: Deep frequency-guided denoising diffusion model for remote sensing image dehazing

Junjie Li, Kaichen Chi, Yue Chang, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Haze removal in remote sensing (RS) images has become increasingly vital due to their capacity to contain essential information for accurate geospatial analysis. Notably, this phenomenon is particularly pronounced in both spatial and spectrum distributions of buildings, complex terrain, and landforms. Inspired by the success of generative models in enhancing details incrementally and suppressing noise, we propose a deep frequency-guided denoising diffusion model for RS imagery dehazing. The pixel-level generative capability of the diffusion model is fully leveraged, and the fast Fourier transform is utilized to extract frequency-domain information. This enables the separate mining of semantic information from RS images in both spatial and spectral domains. Concurrently, the continuity of the image in the frequency domain is ensured without altering the diffusion process, thus achieving detail retention while improving overall clarity. Furthermore, to address the scarcity of physically realistic training data for spatially heterogeneous atmospheric degradation, we construct a Random Haze Distribution Dataset for Remote Sensing dehazing (RHDRS). RHDRS randomly simulates the spatial distribution and thickness of haze, containing 4,500 hazy images along with the corresponding ground truths. Experiments demonstrate that our approach outperforms existing state-of-the-art techniques. The dataset and the code can be accessed at https://github.com/Junjie-LLL/DFG-DDM.**

***Index Terms*—Image dehazing, denoising diffusion models, fast Fourier transform, remote sensing image.**

I. Introduction

In recent years, remote sensing (RS) images have attracted considerable attention from researchers due to their ability to deliver comprehensive surface information and geological distribution features [1]. This capacity renders them invaluable for studies in various fields, including ecological protection, meteorological forecasting, and disaster assessment. However, the acquisition of RS images is affected by real-time atmospheric conditions (*e.g.*, aerosols, cloud cover, and haze). Such poses substantial challenges for fundamental tasks in computer vision, such as semantic segmentation, object detection, and scene understanding.

To address the quality degradation caused by haze, initial efforts primarily utilized physical methods along with deep learning approaches. Physical model-based modalities typically rely on simplified assumptions and imaging principles, such as dark channel prior [5], atmospheric scattering models [6], and haze-line [7] techniques. These techniques are founded on empirical evidence and model the physical principles of haze formation. While physical model-based methods can yield favorable outcomes under specific conditions, their effectiveness diminishes in challenging scenarios. Conversely, deep learning-based modalities [8]–[10] learn intricate patterns and haze fluctuations from large amounts of data. This enables models to discern the distribution and density properties of haze, establishing a mapping relationship between hazy and clear images for effective dehazing. However, these models are more suitable for natural images due to the richer texture information and frequency features in RS images (as shown in Fig. 1). They overlook the intricate patterns, hues, structures, and elements present in RS images, resulting in a loss of image details during the dehazing process.

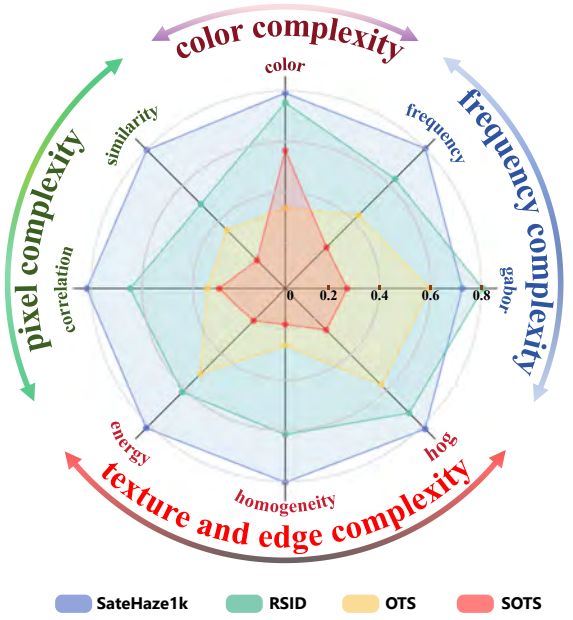

Fig. 1. Schematic illustration of the information sophistication on different types of images, specifically RS image datasets (SateHaze1K [2] and RSID [3]) and natural image datasets (OTS and SOTS [4]). In the radar chart, a total of eight coordinate axes are employed, each represents distinct dataset feature complexity. Specifically, color complexity includes only one axis for color, with value ranges of 0.95-0.98. Pixel complexity encompasses correlation and similarity, where correlation ranges of 0.8-1, and similarity ranges of 3-30. Texture and edge complexity consist of energy, homogeneity, and HOG, with value ranges of 0.01-0.1, 0.06-0.5, and 0.11-0.2, respectively. Frequency complexity includes frequency and Gabor, with ranges of 3300-9000 and 200-250, respectively. To effectively display all features in a single radar chart, we normalize these values to a range of 0-1. A larger area covered by the dataset indicating richer information along the corresponding axis. Clearly, RS images exhibit greater overall information compared to natural images.

As probabilistic diffusion models [11]–[13] achieve better

This work was supported in part by the National Natural Science Foundation of China under Grant 62471394 and U21B2041. (Corresponding author: Qi Wang.)

The authors are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junjieli@mail.nwpu.edu.cn, chikaichen@mail.nwpu.edu.cn, changyue@stu.xjtu.edu.cn, crabwq@gmail.com).

results in various generation tasks, researchers [14]–[16] are endeavoring to apply them to the restoration of images degraded by extreme weather conditions. These diffusion models sequentially add Gaussian noise to data and use a reverse denoising process to generate realistic data, producing better dehazing performance than GANs and autoencoders (AEs). However, the autoencoding convolutional characteristics of diffusion models frequently result in the loss of high-frequency information, which affects the restoration of image details [17]. Furthermore, the structural intricacy of land cover types in RS images (*e.g.*, water bodies, forests, and urban areas) poses additional challenges, resulting in color distortion and diminishing the robustness of the models.

To tackle the aforementioned issues, we suggest a deep frequency-guided denoising diffusion model for RS image dehazing. We propose a data pairing approach to manage size-agnostic images. Given that priors may not be applicable to images with diverse scene data distributions, we do not rely on physical constraints. Instead, we utilize real hazy images combined with frequency domain information to guide the generation results. The integration of high-frequency information obtained from the Fast Fourier Transform into hazy RS images improves the clarity and usability of the images. For the design of the noise prediction network, we first encode the input hazy image information. Following that, we employ a Frequency Guidance Module (FGM) to enable the network to focus on frequency domain information, particularly for extracting high-frequency feature representations. Subsequently, we utilize the UNet architecture to discern the disparities in data distribution patterns between the initial random noise and the final noisy image. Ultimately, a series of Feature Attention Modules (FAM) are employed to facilitate the extraction of information at varying network layers. The frequency-guided diffusion process enables the network to capture both the frequency and spatial characteristics of the image data, ensuring that the final generated results align with the phase and frequency of the real images. Meanwhile, we design a joint loss function to optimize multiple targets while producing images with improved realism. Experiments demonstrate that our method delivers the best performance across various RS dehaze datasets. The contributions of our method can be summarized as follows:

- We design a deep frequency-guided denoising diffusion model that utilizes supervision from the frequency domain to regulate the diffusion process. This innovative approach enables the model to effectively capture the intricate frequency distribution of RS images, significantly enhancing the quality of diffusion-generated results.
- We contribute the RHDRS dataset of randomly distributed haze with varying thicknesses. It provides a platform for scene-constrained supervised learning across diverse environmental conditions. As a result, this dataset fills a critical gap in benchmarking and promotes advances in RS dehazing research.
- Our method achieves leading performance in the quality of RS image dehazing, providing new improvements and insights for tackling challenging real-world scenarios.

II. Related Work

A. Remote Sensing Dehazing

For image dehazing, researchers initially explore natural images [18]–[20]. Berman *et al*. [21] proposed that a dense cluster of colors could effectively represent the hues of clear scenes, leading to the introduction of a haze-line prior for enhanced dehazing. Li *et al*. [22] reconstructed the atmospheric scattering model to produce restored images. These mothods achieve satisfactory results in dehazing natural images.

Due to the presence of surface features such as color and structure in remote sensing (RS) images, natural image dehazing methods often fail to effectively restore these images. As a result, researchers have conducted related explorations specifically for dehazing RS images. Song *et al*. [23] combined the transformer and U-Net frameworks to restore images through multiple layers of feature extraction and representation. Chi *et al*. [3] applied a gradient-guided strategy to the Swin-Transformer, using three predictors to learn haze parameters. However, these methods overlook the frequency domain characteristics of RS images, culminating in a certain degree of frequency information loss. In contrast, our approach ensures the continuity of the restored image information in both the spatial and frequency domains by leveraging frequency domain information guidance.

B. Fast Fourier Transform

In recent years, frequency domain learning has attracted increasing research attention in pursuit of effective spectral features. Researchers have utilized Fourier transforms to extract frequency domain information and successfully applied it to various tasks, such as classification [24], denoising [25]–[27], and dehazing [28]–[30]. Cai *et al*. [28] proposed a new frequency domain image translation framework that enhances the image generation process using frequency information. Fu *et al.* [29] employed frequency information through a two-dimensional discrete wavelet transform to train a generative adversarial network (GAN) that directs the image dehazing process. Nevertheless, frequency domain approaches frequently overlook variations in local features. Additionally, frequency domain processing is susceptible to noise, which hinders its adaptability to different haze levels in various regions. We leverage frequency domain information to steer the diffusion process, enhancing noise suppression and refining local image details while retaining frequency information.

C. Diffusion-based Generative Models

Diffusion-based generative models [12] utilize parameterized Markov chains to enhance the lower variational bound of the likelihood function. They progressively corrupt images by adding noise and iteratively denoise from a noise distribution to restore clear images. This approach appears effective in several image processing applications, including image synthesis [31], deblurring [32], and dehazing [14]–[16]. Dhariwal *et al.* [31] enhanced the quality of image generation by combining conditional control with an upsampling strategy to guide image generation. Luo *et al.* [14] optimized the

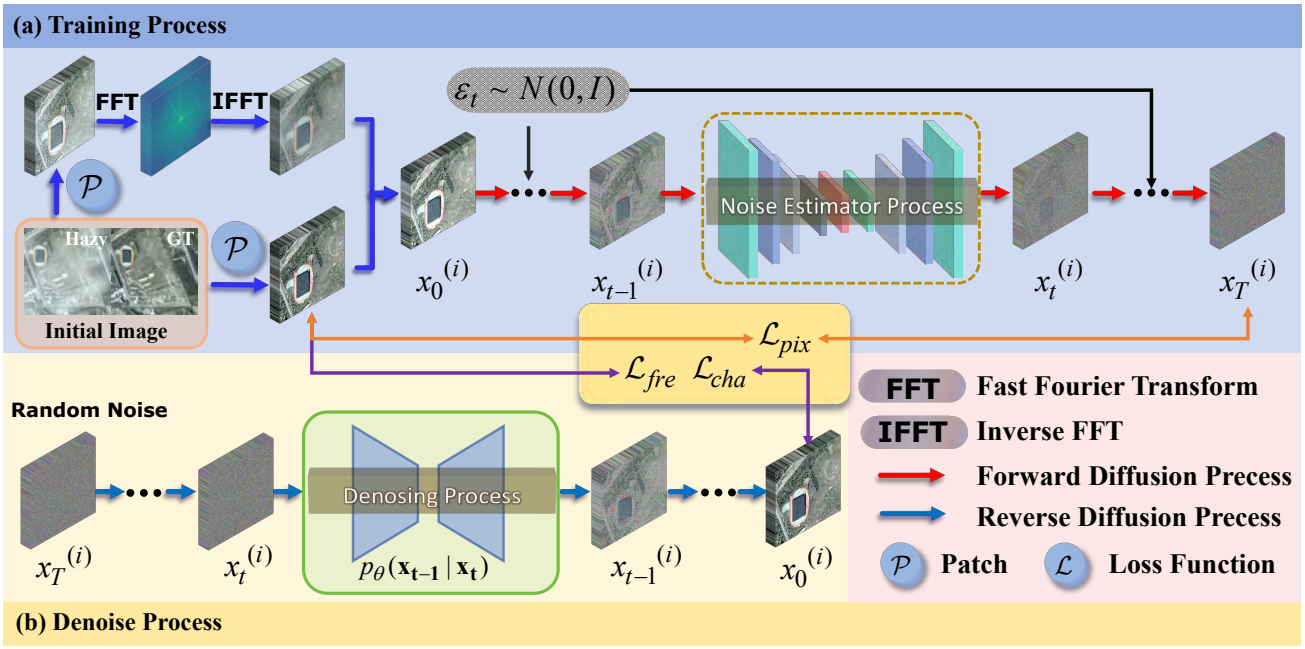

Fig. 2. Overview of the architecture of our proposed DFG-DDM. (a) Training process of the diffusion model, where hazy images serve as control conditions. In the information flow, hazy images undergo Fourier transform to obtain high-frequency dominated frequency domain information as conditional control. This information is subsequently combined with ground truth (GT) images to serve as the input for the training process, which enables the prediction of noise through the Noise Estimator Process. (b) Denoise Process. The input random Gaussion noise is processed through a noise prediction network for denoising, ultimately restoring it to a clear image. The denoising process includes the Frequency Guidance Module (FGM) and the Feature Attention Module (FAM). These modules collaborate within the network to extract features, as illustrated in Fig. 3. The network is optimized using joint loss functions: pixel loss ($\mathcal{L}_{pix}$), frequency loss ($\mathcal{L}_{fre}$), and Charboiner loss ($\mathcal{L}_{cha}$).

diffusion model with a U-Net based latent diffusion strategy and fine-tuning hyperparameters to enhance restoration quality. Although diffusion model-based approaches effectively learn spatial information from images, they typically struggle to capture high-frequency features, particularly in RS images. This limitation hampers the consistency of frequency domain information in the restored images. The effective integration of frequency domain information with the diffusion model is a critical consideration.

III. METHOD

A. motivation

The denoising diffusion model prioritizes spatial information and pixel data distribution, which affects the representation of image frequency components. These spectral features are particularly critical for remote sensing (RS) image restoration, as they preserve essential geological structure, surface textures, and color fidelity. Extracting information from both spatial and spectral domains simultaneously facilitates enhanced verisimilitude in RS data reconstruction. Considering the strengths of diffusion models in image restoration, incorporating frequency domain features into the diffusion process has the potential to yield excellent RS dehazing results.

B. Denoising Diffusion Probabilistic Models

The diffusion model is a generative approach implemented through sequential Markov chains. Its core consists of the training process and the reverse inference process. Fig. 2 illustrates the architecture of our diffusion model.

Forward Diffusion Process. In the forward diffusion process, diffusion model adds noise to transform the input image $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ into noisy image $\mathbf{x}_T \sim \mathcal{N}(0,1)$ through T iterations. The process of adding noise can be viewed as a Markov chain. Due to the propagation property of the Markov chain, the data at any time-step $t \in [0,T]$ can be obtained from the data at time-step $t-1$. Therefore, we can obtain the following noise data distribution as:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathcal{Z}\right), \tag{1}$$

where β_t is the scale factor to control the amount of noise added at time-step t. $\mathcal{Z}$ is the noise sampled from a standard normal distribution. Through the parameter renormalization trick, the Gaussian noise distribution at t-th step can be simplified to $q\left(x_t \mid x_0\right)$:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)\mathcal{Z}), \tag{2}$$

where $\overline{\alpha}_t = \prod_{i=1}^{T} \alpha_i$. When the number of iterations t becomes sufficiently large, this distribution $q(x_t|x_0)$ gradually approaches a normal distribution $\mathcal{N}(0,\mathcal{Z})$. At this point, the diffusion model can be considered to have completed the forward diffusion process.

Reverse Diffusion Process.

The reverse diffusion process removes noise through T iterations to restore the original distribution from the Gaussian noisy data. Similar to the forward diffusion process, given the condition x_0, diffusion model first samples x_T from a Gaussian distribution and then gradually denoises it until a high-quality output is obtained:

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}\left(\mathbf{x}_t, \mathbf{x}_0\right), \tilde{\beta}_t \mathcal{Z}\right), \tag{3}$$

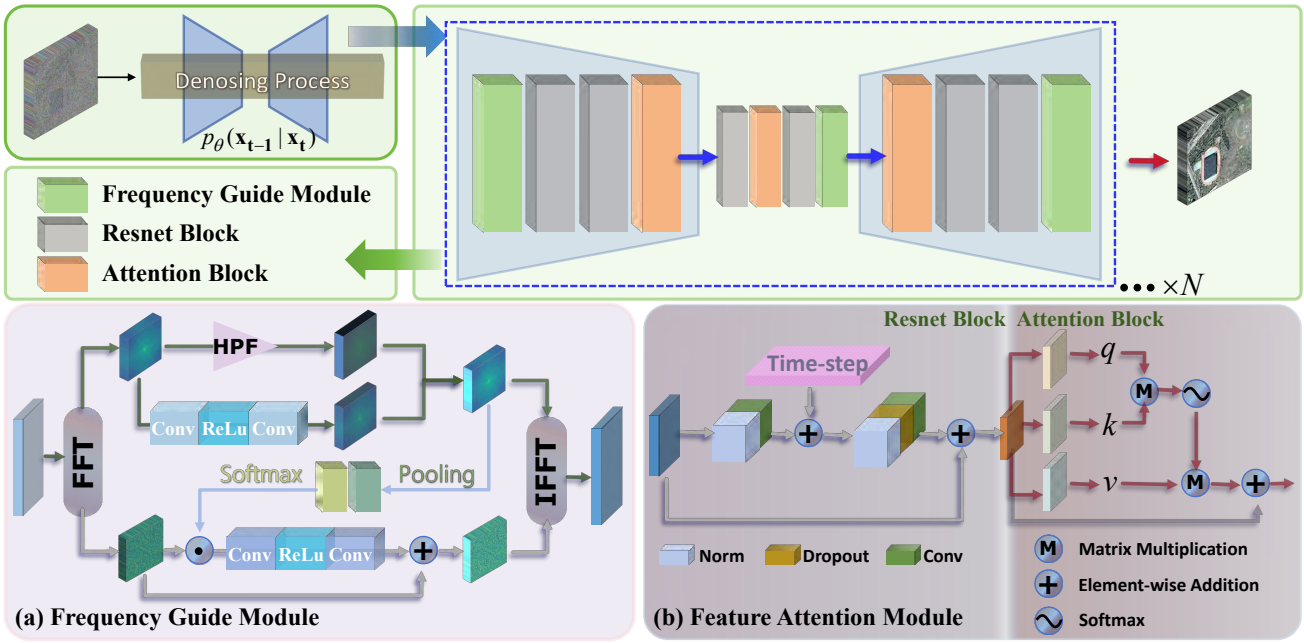

Fig. 3. The implementation details of the denoising network, which mainly consists of the Frequency Guide Module, ResNet Block, and Attention Block. (a) The Frequency Guide Module utilizes the input information to obtain amplitude and phase through FFT. After a series of information extraction and fusion processes, the frequency domain information is guided through IFFT, with HPF representing a high-pass filter. (b) The Feature Attention Module includes both the ResNet Block and the Attention Block, enhancing the understanding of image information through the design of residual convolutions and attention mechanisms.

Algorithm 1 Inference Process of FDG-DDM.

Require: noisy image x_T, contional image I'
Ensure: restore image $\tilde{x}_0$
1: Randomly sample a noise x_T from $\mathcal{N}(\mathbf{0}, \mathcal{Z})$
2: **for** $t = T, \ldots, 1$ **do**
3: Concat x_T and I' into the network
4: Calculate $\epsilon_t \leftarrow \epsilon_\theta(x_t, I', t)$
5: $x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\cdot\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, I', t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \epsilon_\theta(\mathbf{x}_t, I', t)$,
6: **end for**
7: **return** $\tilde{x}_0$

where $\tilde{\mu}(x_t, x_0)$ is the mean of the probability distribution and $\tilde{\beta}_t$ is the variance:

$$\tilde{\mu}(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right), \tag{4}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t), \tag{5}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathcal{Z})$ represents the noise in x_t at time-step t.

However, in image generation, estimating $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$ is challenging due to the inherent uncertainty of the entire dataset. We need to learn a model $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ to approximate it, which is uncontrol of x_0:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathcal{Z}), \tag{6}$$

Notably, $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ represents the model estimated distribution, while $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$ is the ground-truth data distribution from the reverse diffusion process. To ensure that the model generates accurate results, these two distributions need to be as close as possible.

For $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, let $\sigma_t^2 = 1 - \alpha_t$, then we need to predict $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ as closely as possible to $\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0)$ to generate high quality results. This process is achieved using a denoising network. Equivalently, this implies adopting a network $\epsilon_\theta(x_t, t)$ to estimate ϵ_t. To train $\epsilon_\theta(x_t, t)$, the diffusion model adds noise to the image based on the known x_0 through a series of time steps, using the method described in Equation 2. Then, by refining the adjusted simplified goal θ of $\epsilon_\theta(x_t, t)$ to better approximate ϵ_t, this process ultimately achieves noise prediction.

C. Frequency domain information-guided diffusion process

The frequency domain information is crucial for the recovery of RS haze images. The spectral decomposition based on Fourier transform facilitates the extraction of critical image attributes, addressing the information loss associated with spatial-focused restoration approaches. Its high-frequency components are fused with the original data, collectively serving as control conditions for the diffusion model.

Specifically, for a two-dimensional image with a resolution of $m \times n$, its two-dimensional discrete Fourier transform (DFT) as follows:

$$F(a, b) = \sum_{r=0}^{R-1}\sum_{c=0}^{C-1} I(r, c) e^{-i2\pi\left(\frac{ar}{R} + \frac{bc}{C}\right)}, \tag{7}$$

where $F(a, b)$ is the component of the image at frequency (a, b), $I(r, c)$ is pixel values of the image at position (r, c). To efficiently compute the DFT, we use the Fast Fourier

Transform (FFT) to extract frequency domain information F from the image.

To filter out the low-frequency information of haze images and retain distinct high-frequency components, we create a high-pass filter H with a threshold of θ:

$$H(a,b) = \begin{cases} 1, & \text{if } D(a,b) > \theta \\ 0, & \text{if } D(a,b) \leq \theta \end{cases}, \tag{8}$$

where $D(a,b)$ is the distance to the frequency domain origin. The threshold θ is a dynamic parameter defined as a percentage of the frequency information content of the data. Frequencies that exceed this threshold are designated as high-frequency information, while others are ignored. By adjusting θ, high-frequency components are adaptively filtered across datasets, achieving a balanced representation while effectively suppressing noise based on varying noise distributions.

According to Equation 9, we apply the high-pass filter H to the frequency domain image, performing element-wise multiplication with the frequency components $F(a,b)$, where G is the extracted high-frequency information.

$$G(a,b) = F(a,b) \cdot H(a,b), \tag{9}$$

In the process of dehazing RS images, we introduce additional high-frequency information to enhance the frequency domain data, merging it with the original image according to the following formula:

$$I_h = \mu G(a,b) + (1-\mu)F(a,b), \tag{10}$$

where I_h is the frequency domain feature enhanced for high-frequency information, μ is the hyperparameter.

To align this portion of frequency information with the dimensions of the input image x_0, we utilize the Inverse Fast Fourier Transform (IFFT) to convert the frequency domain image back to the spatial domain, defined as:

$$I' = \frac{1}{RC} \sum_{a=0}^{R-1} \sum_{b=0}^{C-1} I_h(a,b) e^{j2\pi\left(\frac{ra}{R} + \frac{cb}{C}\right)}, \tag{11}$$

where I' is the integrated frequency domain image information.

To enhance the spatial and frequency domain features of the generated images, we design a diffusion process guided by the fused image information I'. It should be noted that incorporating I' into the regulation of the probability distribution $p_\theta(x_{0:T}|I')$ does not affect the diffusion process $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$. Similar to the reverse process of the diffusion model (Equation 3), we predict the unknown quantity $\epsilon_\theta\left(x_t, I', t\right)$ through a neural network:

$$\epsilon_\theta\left(x_t, I', t\right) \sim \epsilon_t. \tag{12}$$

Specifically, we concatenate the control image and the original image along the dimensions and use the result as input to the network. After network prediction, we can calculate x_{t-1} through x_t and continue this process until we finally obtain the final clear image $\tilde{x}_0$ that meets our quality standards. The details of the inference process of the frequency domain-guided denoising diffusion model (FDG-DDM) are shown in Algorithm 1.

D. Noise Prediction Network Framework

We use an improved UNet as our backbone network to predict the noise distribution, as shown in Fig. 3. In the frequency domain, the degradation characteristics of two-dimensional hazy images are primarily reflected in the amplitude spectrum. Consequently, we design the Frequency Guidance Module (FGM) and Feature Attention Module (FAM) to jointly guide the predicted noise on both spatial and spectrum domain information of images. The FGM primarily focuses on extracting frequency domain information during the diffusion process. It emphasizes high-frequency components in hazy images to enhance the detailed features of RS images. In contrast, the FAM targets the spatial characteristics of images by learning spatial feature distributions through residual blocks and attention mechanisms. Additionally, the FAM is guided by the FGM, enabling the network to integrate frequency and spatial domain information. This collaborative framework improves the accuracy of noise prediction and enhances image restoration performance.

Frequency Guidance Module. Haze causes nonuniform illumination attenuation, mainly degrading low-frequency features while preserving phase spectrum. To effectively suppress haze and adapt to different environments, we focus on high-frequency information, ensuring better preservation of image details and edges.

According to Equation 7, base on the frequency domain information $\mathcal{F}(x)$, we can obtain the amplitude $\mathcal{A}(x)$ and the phase $\mathcal{P}(x)$:

$$\mathcal{A}(x)(i,j) = \sqrt{[\mathcal{R}^2(x)(i,j) + \mathcal{I}^2(x)(i,j)]}, \tag{13}$$

$$\mathcal{P}(x)(i,j) = \arctan\left[\mathcal{I}(x)(i,j)/\mathcal{R}(x)(i,j)\right], \tag{14}$$

where $\mathcal{R}(x)$ and $\mathcal{I}(x)$ are the real and imaginary parts of $\mathcal{F}(x)$, respectively.

To make the network focus more on the high-frequency information in the spectrum domain, we also use a high-pass filter (Equation 8) to extract features in the convolution process, as follows:

$$\mathcal{A}_h = Conv_{1\times1}(\mathcal{A}(i,j) \cdot \mathcal{H}(i,j)), \tag{15}$$

where $\mathcal{A}_h$ is the enhanced high-frequency information. Combining the concept of residual networks, we use $\mathcal{A}_{res}$ and $\mathcal{A}_h$ to jointly guide the optimization of the phase:

$$\mathcal{A}_{res}(x)(i,j) = Conv_{1\times1}(\mathcal{A}'(x)(i,j)) - \mathcal{A}(x)(i,j), \tag{16}$$

$$\mathcal{A}'(x)(i,j) = \mathcal{A}_{res}(x)(i,j) + \mathcal{A}_h, \tag{17}$$

Through the combination of $\mathcal{A}_{res}$, the refined frequency output $\mathcal{A}'$ is obtained. After obtaining amplitude $\mathcal{P}'$ through information aggregation, we convert $\mathcal{P}'$ and $\mathcal{A}'$ into the real and imaginary parts of the newly recovered frequency domain by:

$$\begin{aligned} \mathcal{R}^{'}(x)(i,j) &= \mathcal{A}^{'}(x)(i,j)\cos\mathcal{P}^{'}(x)(i,j), \\ \mathcal{I}^{'}(x)(i,j) &= \mathcal{A}^{'}(x)(i,j)\sin\mathcal{P}^{'}(x)(i,j), \end{aligned} \tag{18}$$

Finally, we use the IFFT to recover the spatial domain representation of the information:

$$\mathcal{Y}_{fre} = irfft2(\mathcal{R}' + \mathcal{I}'j), \tag{19}$$

where $\mathcal{Y}_{fre}$ is the output of frequency guidance module.

Feature Attention Module. To capture global image information during the diffusion process and effectively extract details and key features, we design a Feature Attention Module (FAM). The FAM leverages the features of the UNet backbone and consists of multiple residual blocks along with an attention mechanism block. For each residual block, two convolutional processes are utilized for information extraction. These processes incorporate the sinusoidal positional encoding Emb_t for each sampling step t, as shown in the following equation:

$$x_{temp} = Conv(\delta(N(x))) + Emb_t, \tag{20}$$

$$x_{end} = Conv(D(\delta(N(x_{temp})))) + x, \tag{21}$$

where x is the input feature obtained from the FGM module, δ represents a nonlinear activation function, N and D represent the normalization and dropout operations. After each residual block, we use downsampling to aggregate image information by setting the convolution stride to 2, effectively reducing the resolution of the feature map.

When the image resolution is downsampled to 16×16, we apply an attention block to the channel image. We mainly use 2D convolution operations to implement the attention mechanism. First, we apply 2D convolution kernels independently to the input features x' to extract q, k, and v as query, key, and value features, respectively. The attention weight matrix w is computed by reshaping q and k into three dimensions and performing over the channel c:

$$w[b,i,j] = \sum_c q[b,i,c] \cdot k[b,c,j], \tag{22}$$

where $w \in \mathbb{R}^{b\times(h\cdot w)\times(h\cdot w)}$. Subsequently, we apply the softmax function along the last dimension of w to obtain the normalized weights w'. Finally, we obtain the output features y through convolution:

$$y = Conv(w' \otimes v) + x'. \tag{23}$$

E. loss function

During the network training process to predict noise, we design a joint loss function to optimize network parameters. Similar to [15], We use the L2 norm of the predicted noise $\hat{I}_i$ and the real noise I_i as the pixel loss function $\mathcal{L}_{pix}$ to reduce training bias for the diffusion process:

$$\mathcal{L}_{pix} = \frac{1}{N}\sum_{i=1}^{N}(I_i - \hat{I}_i)^2, \tag{24}$$

We use the frequency loss [30] $\mathcal{L}_{fre}$ to maintain the consistency of frequency domain information, which consists of $\mathcal{A}$ and $\mathcal{P}$:

$$\mathcal{L}_{fre} = \frac{2}{MN}\sum_{m=0}^{M/2-1}\sum_{n=0}^{N-1}\left(\||\Delta_{\mathcal{A}}|_{m,n} + |\Delta_{\mathcal{P}}|_{m,n}\|_1\right), \tag{25}$$

where Δ is the difference between the corresponding output and the initial input.

Additionally, to enhance the capture of image edge information, we introduce the Charbonnier loss [33] $\mathcal{L}_{cha}$ to refine the differences between the predicted noise and the real noise:

$$\mathcal{L}_{cha} = \frac{1}{hwc}\sum_{x,y,z}\sqrt{(\hat{I}_{x,y,z} - I_{x,y,z})^2 + \xi^2}, \tag{26}$$

where h, w, c represent the height, width, and channels of the image, respectively. ξ is a parameter that enhances numerical stability.

Finally, to balance the effects of each component of the loss function, following [30], we set the hyperparameters $\mu = 10$ and $\sigma = 50$. The joint loss function $\mathcal{L}$ is computed by adding the individual loss components.

$$\mathcal{L} = \mathcal{L}_{pix} + \mu\mathcal{L}_{fre} + \sigma\mathcal{L}_{cha}. \tag{27}$$

Our proposed joint loss function combines pixel loss, frequency loss, and Charbonnier loss. Building upon pixel loss, frequency loss ensures the preservation of frequency and texture, while Charbonnier loss enhances robustness to outliers and noise. This approach captures both local and global features, improving adaptability to complex hazy environments. Consequently, the performance of dehazing is significantly improved, yielding images that are both lifelike and aesthetically pleasing.

IV. EXPERIMENTS

A. Data Generation

The thickness and spatial distribution of haze are influenced by meteorological conditions, topography, and human ativities, resulting in notable randomness and unevenness. To solve the lack of haze variability in RS images, we construct a Random Haze Distribution dataset for Remote Sensing images (RHDRS). The RHDRS dataset is collected from the RSICD [34], RSOD [35] and NWPU-RESISC [36] datasets. It contains 21 scenario categories, including airplanes, airports, bridges, bushes, circular farmland, commercial areas, dense residential areas, deserts, forests, highways, lakes, overpasses, oil storage tanks, meadows, terraces, harbors, parking lots, mountains, intersections, churches, and beaches. Among these categories, 280 images are available for each scene of airport, commercial area, dense residential area, round farmland, overpass, and intersection. Additionally, 220 images are available for each scene of airplane, harbor, oil storage tank, forest, church, and parking lot. There are 200 images for each of the lake, beach, bush, and grass scenes. The remaining scenes range between 100 and 150 images per category. Regarding the diversity of haze distribution in the RHDRS dataset, 30% of the images feature light haze, 40% have moderate haze, and 30% exhibit heavy haze. From the perspective of scene density, the high-density scenes account for 40%, including commercial areas, industrial areas, and other scenes. The remaining scenes comprise more variations and less correlation, spanning overpasses, farmland, forests, oceans, oil storage tanks, and mountains, which collectively showcase rich and intricate image characteristics. The RHDRS dataset consists of 4,500 images pairs with a resolution of 128×128. We randomly split the dataset into 4,000 image pairs for training

Fig. 4. Some samples from the RHDRS dataset, illustrating diverse scenes such as airplanes, bridges, harbors, highways, and other landscapes. The upper portion of each image displays scenes characterized by a light synthetic haze, while the lower portion exhibits scenes with a denser haze. Both the thickness and spatial distribution of the haze are randomized.

and 500 pairs for testing. Images of different scenarios and haze distribution are involved in both the training and test sets. As these collected images do not include reference hazy images, we apply a haze synthesis algorithm to construct annotated pairs for the RHDRS dataset, as shown in Fig. 4.

According to the haze image synthesis theory proposed by [37], given a clear image $J(x, y)$, the hazy image can be synthesized using medium transmission map $t(x, y)$ and global atmospheric light A:

$$I(x, y) = J(x, y)t(x, y) \oplus As(x, y), \tag{28}$$

where (x, y) represents the pixel location, $s(x, y)$ represents the haze thickness delineates the organized haziness. To control the thickness of haze, we set the ambient light intensity $A \in [0.6, 1]$, the medium transmission $t \in [0.2, 0.8]$, and the haze thickness $s \in [0.2, 0.9]$. For the random spatial distribution, we generate a two-dimensional random noise map $N(x, y)$ within the range [0, 1] using the Gaussian distribution. By adjusting the spatial distribution, we generate RS haze images with variable haze density and spatial distribution:

$$I(x, y) = J(x, y) \cdot t(x, y) \oplus A \cdot s(x, y) \cdot N(x, y), \tag{29}$$

where $N(x, y)$ is the noise value corresponding to each pixel. This value controls haze intensity distribution, creating spatial randomness for a more natural appearance.

B. Experiment Settings

In this section, we elaborate on the implementation details, datasets, compared methods, and evaluation metrics.

Implementation Details. Our proposed DFG-DDM network is constructed on the UNet architecture, integrating deep Fourier convolution and attention mechanisms during both downsampling and upsampling. The channel numbers of the network are set to $(128, 128, 256, 256, 512, 512)$, with two ResNet blocks added to each layer to enhance its capacity. The channel number for the attention module is set to 16, and a deep Fourier convolution block is performed before each Unet sampling process begins. To address the computational complexity concerns, we also introduce a lightweight variant of the DFG-DDM network, denoted as DFG-DDM-Lite. This variant reduces the channel numbers to $(128, 128, 256, 512)$ and employs a single ResNet block per layer, while maintaining the attention module and deep Fourier convolution block. The adjusted network depth effectively captures essential frequency-domain information, while a single residual block robustly supports detail reconstruction, ensuring high-quality dehazing performance. We train and test model on a single NVIDIA GeForce RTX 3090 GPU, performing 2,000,000 iterations for each dataset with a learning rate of 1×10^{-4}. To address the variability in dataset sizes, we adopt a patch strategy. The patch size is configured to 64×64, and the batch size is set to 10. We use the Adam optimizer and set the exponential smoothing parameter to 0.999 to enhance stability during the training process.

Datasets. We test the effectiveness of our model on several publicly available datasets: RSID [3], RSHaze [23], SateHaze1k [2], and RHDRS. The SateHaze1K dataset consists of three subdatasets: thin, moderate, and thick. Each subdataset comprises 400 pairs of synthetic RS hazy images sized 512×512, with 320 images allocated for training, 35 for validation, and the remaining 45 for testing. We experiment on these three subsets following the authors' settings. Meanwhile, following the configuration of RSID and RSHaze, we combine them to SateHaze1k-all, selecting 1,100 images for training and 100 for testing. The RSHaze dataset includes 54,000 image pairs with a size of 512×512. Following predefined splits of the authors, we use 51,300 pairs for training and 2,700 pairs for testing. The RSID dataset contains 1,000 pairs

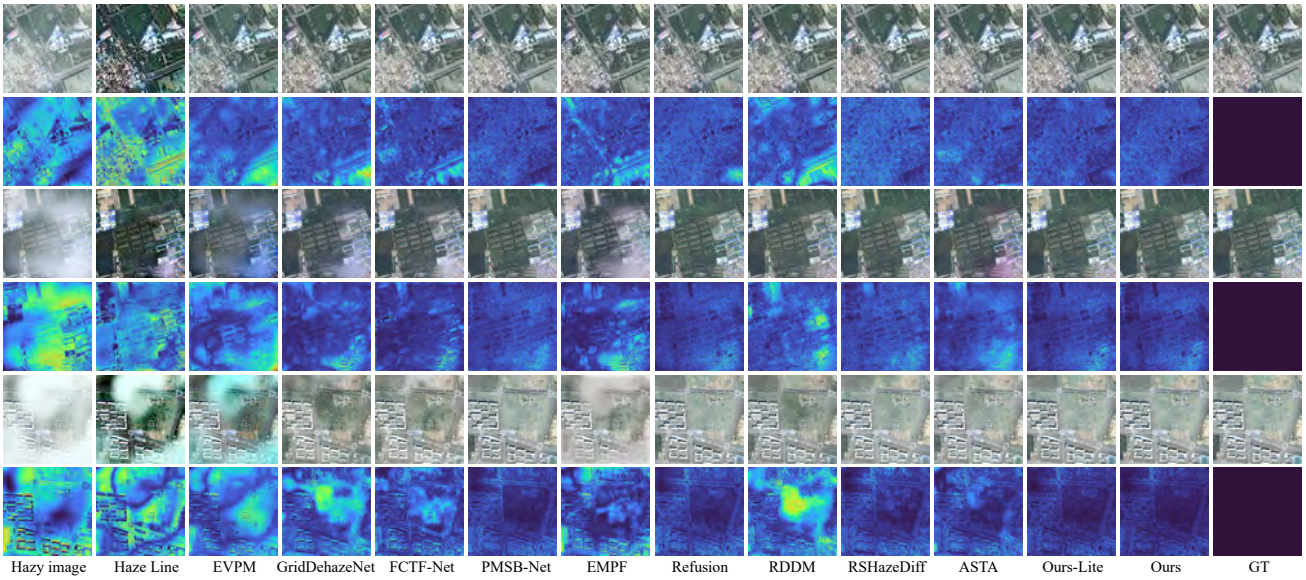

Fig. 5. The dehazing results of the SateHaze1k dataset are presented here, accompained by corresponding error for enhanced visualization. The first two rows show results for thin haze, the middle two rows for moderate haze, and the last two rows for thick haze. In the error maps, lower pixel values signify a closer alignment with the ground truth.

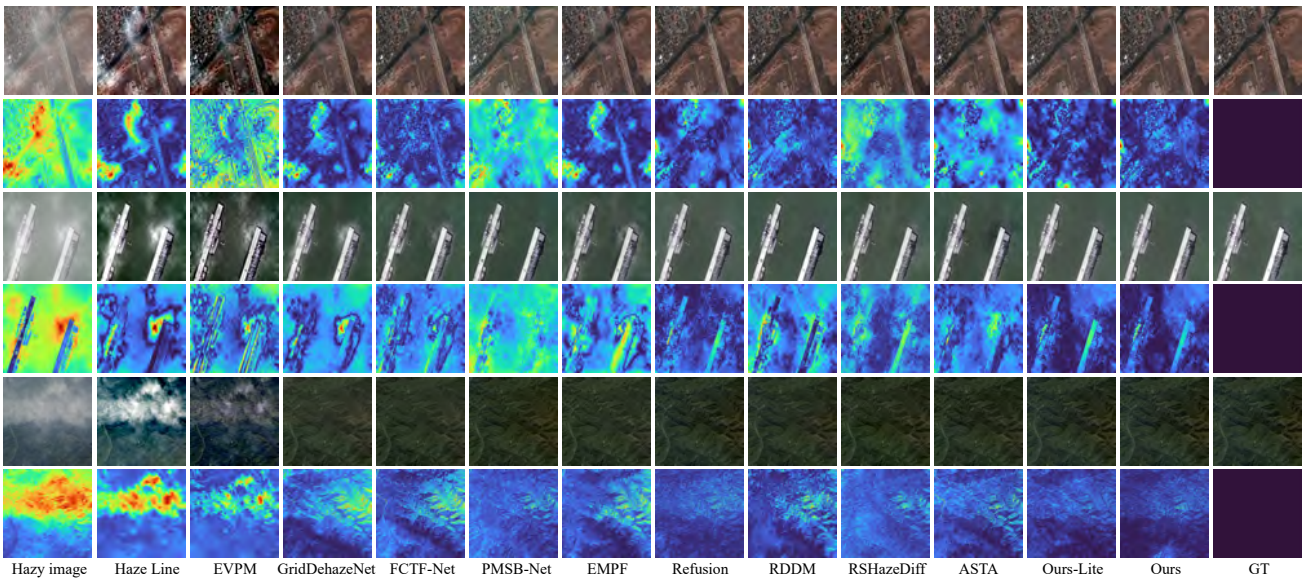

Fig. 6. The visual results on the RSID (top four rows) and RSHaze (bottom two rows) datasets demonstrate that our method delivers superior performance under these two synthesized haze distribution datasets.

of images, each sized 128×128. We randomly select 900 pairs for training and the leftover images for test. For the RHDRS dataset, it contains a total of 4,500 image pairs with a size of 128×128. We classify the haze levels and randomly select images from each level, ultimately forming a training set of 4,000 pairs and a test set of 500 pairs.

Compared Methods. We evaluate our DFG-DDM against the following state-of-the-art methods, including **traditional methods:** Haze-line [7], ROP [38], EVPM [39]. **Natural image dehazing methods:** FFA-Net [40], 4KDehazing [41], FSDGN [30], GridDehazeNet [42], DEA-Net [43], Refusion [14], RDDM [12]. **RS image dehazing methods:** Dehazeoformer [23], TrinityNet [3], EMPF-Net [44], ASTA [45], FCTF-Net [46], PMSB-Net [47], RSHazeDiff [16].

Evaluation metrics. For these datasets, we perform full reference evaluations using PSNR [15], SSIM [15], and LPIPS [48] to assess the dehazing performance of the model. PSNR and SSIM measure pixel-level fidelity and structural coherence, respectively. Higher scores reflects superior quality. LPIPS evaluates perceptual similarity, rendering it particularly pertinent for assessing the utility of RS imagery in subsequent tasks. Lower values indicate greater consistency between dehazed and ground-truth images.

C. Visual Comparisons

We first present a comparative analysis of the StateHaze1k dataset. As illustrated in Fig. 5, the images range from low to high haze intensity levels, offering a comprehensive represen-

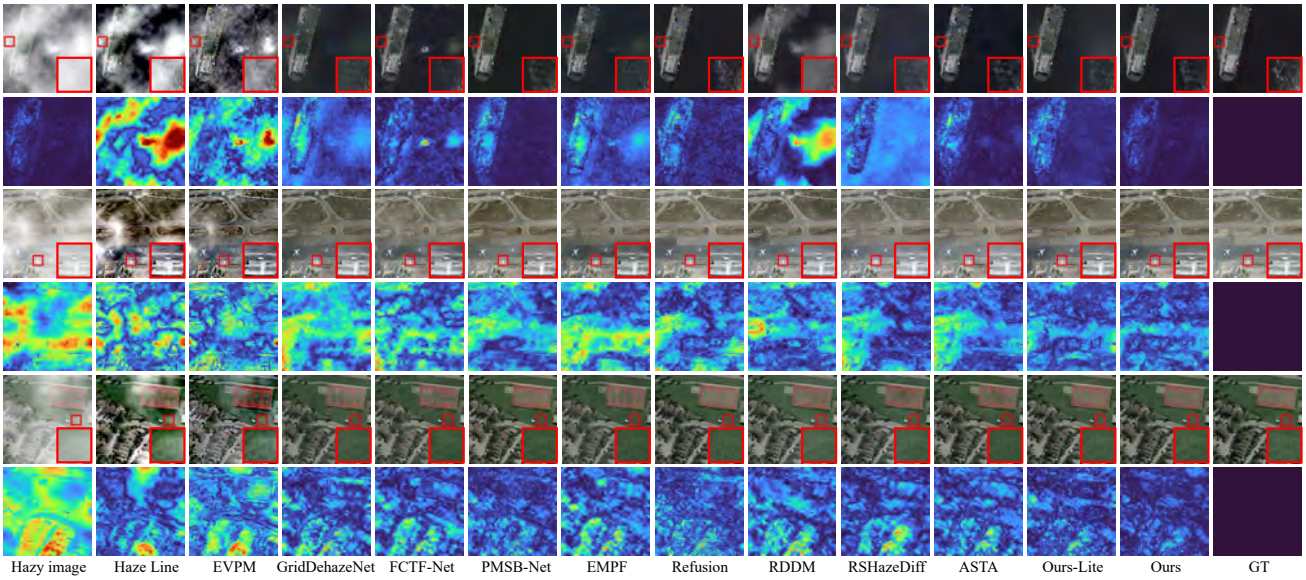

Fig. 7. The RHDRS dataset offers realistic haze with varying densities and random distributions across diverse environments, providing a solid foundation for training dehazing algorithms. We showcase the results of dehazing under various haze conditions.

TABLE I
THE AVERAGE PSNR (DB), SSIM, LPIPS, AND THE COMPUTATIONAL COMPLEXITY ON **RSID**, **RSHAZE** AND **RHDRS** DATASETS. THE BEST SCORE IS IN RED AND THE SECOND BEST IS IN BLUE. OUR METHOD ACHIEVES SUPERIOR RESULTS ACROSS ALL DATASETS.

Methods	Publication	RSID			RSHaze			RHDRS			Computational Complexity	
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	# Params ↓	FLOPs ↓
haze-line [7]	TPAMI'20	16.1062	0.7678	0.1603	13.0602	0.5759	0.3351	12.5834	0.5692	0.2776	-	-
EVPM [39]	INS'22	16.2948	0.7014	0.1691	16.9379	0.5622	0.2880	15.8803	0.6328	0.2223	-	-
ROP [38]	TPAMI'23	12.5183	0.6115	0.2818	8.3104	0.3805	0.4553	9.6650	0.5022	0.3752	-	-
GridDehazeNet [42]	ICCV'19	20.7516	0.8836	0.0665	30.3892	0.9107	0.0672	23.6355	0.8683	0.0746	0.96M	38B
FFA-Net [40]	AAAI'20	26.8578	0.9468	0.0251	37.5184	0.9503	0.0382	27.0309	0.8949	0.0621	4.46M	144B
FCTF-Net [46]	GRSL'21	22.7012	0.8815	0.0750	31.7223	0.9077	0.0785	23.7278	0.8543	0.0901	0.16M	40B
FSDGN [30]	ECCV'22	27.6927	0.9561	0.0187	35.4054	0.9459	0.0419	27.6172	0.9011	0.0563	2.73M	79B
EMPF-Net [44]	TGRS'23	21.2157	0.8908	0.0635	34.1121	0.9322	0.0559	23.9917	0.8688	0.0762	0.43M	12B
PSMB-Net [47]	TGRS'23	25.2069	0.9447	0.0285	34.4964	0.9445	0.0380	25.4464	0.8930	0.0596	13.6M	98B
DehazeFormer [23]	TIP'23	26.1513	0.9417	0.0283	37.5440	0.9475	0.0422	26.7499	0.8992	0.0602	2.52M	26B
ASTA [45]	GRSL'24	24.9798	0.9160	0.0435	34.3136	0.9198	0.0741	26.2774	0.8870	0.0683	4.89M	68B
DEA-Net [43]	TIP'24	26.1590	0.9448	0.0221	37.5868	0.9503	0.0275	26.7236	0.8959	0.0532	7.79M	20B
4KDehazing [41]	CVPR'21	23.9095	0.9312	0.0389	32.3704	0.9322	0.0543	24.3458	0.8759	0.0713	34.55M	104B
Trinity-Net [3]	TGRS'23	24.0516	0.8993	0.0537	33.9013	0.9304	0.0415	24.9373	0.8926	0.0678	33.31M	99B
Refusion [14]	CVPR'23	27.3763	0.9370	0.0219	36.6953	0.9207	0.0313	27.3140	0.8645	0.0507	131.4M	64B
RDDM [12]	CVPR'24	25.4104	0.9450	0.0234	32.0684	0.9093	0.0578	24.1198	0.8697	0.0721	46.3M	10B
RSHazeDiff [16]	TITS'24	23.7694	0.8695	0.0505	32.3339	0.8831	0.1015	24.8747	0.8806	0.0630	110.2M	42B
DFG-DDM-Lite	–	27.7136	0.9573	0.0192	38.2219	0.9521	0.0249	27.6244	0.9006	0.0512	33.7M	54B
DFG-DDM	–	27.8517	0.9590	0.0103	38.8261	0.9549	0.0230	27.8044	0.9028	0.0463	113M	68B

tation of the variability in RS imagery. Traditional methods fail to restore the expected clarity (*e.g.* haze-line [7] and EVPM [39]). For deep learning-based methods, GridDehazeNet [42] and RDDM [12] lack atmospheric model support, leading to a loss of ground object details. RSHazeDiff [16] and ASTA [45] lack frequency information and finer texture details. In contrast, our method excels at restoring ground object details and preserving subtle color and frequency features.

Next, we compare methods on the RSID and RSHaze datasets, which specifically focus on urban, agricultural, and coastal areas, as depicted in Fig. 6. RS dehazing methods (*e.g.*, PMSB-Net [47], and RSHazeDiff [16]) generally perform better in detail preservation of surface buildings. But they still introduce color inconsistencies and deviations after dehazing. Our method demonstrates superior performance in detail recovery and color consistency, particularly in preserving edges and textures of ground features.

Finally, we present the visual effects of our contributed RHDRS dataset in Fig. 7. Conventional natural image dehazing methods often fail to achieve optimal visual performance, as they tend to excessively remove non-haze-related image features. Our method excels in haze removal, preserving original image details with minimal content degradation. This showcases the adaptability and robustness across diverse RS scenarios.

D. Quantitative Comparisons

To ensure fair and reliable quantitative comparisons, we employ the original authors' source code to retrain the competing methods on a standardized training set. Additionally, we report comparisons between our model and others with similar parameters. We report PSNR, SSIM, and LPIPS scores for each method in Table I and Table II. On the RSID,

TABLE II
THE AVERAGE PSNR (DB), SSIM, LPIPS ON THE DIVIDED **SATEHAZE1K** DATASETS. THE BEST SCORE IS IN RED AND THE SECOND BEST IS IN BLUE. OUR METHOD ACHIEVES SUPERIOR RESULTS ACROSS THE DIFFERENT SUBSET DATASETS.

Methods	Publication	SateHaze1k-thin			SateHaze1k-moderate			SateHaze1k-thick			SateHaze1k-all		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
haze-line [7]	TPAMI'20	14.6972	0.7178	0.1516	14.6472	0.7650	0.1969	12.9452	0.4682	0.3496	14.4647	0.7020	0.2034
EVPM [39]	INS'22	16.0498	0.7809	0.1619	17.5223	0.7961	0.1316	13.2915	0.5186	0.3380	17.8292	0.8061	0.1284
ROP [38]	TPAMI'23	13.1833	0.6815	0.2211	14.4952	0.7119	0.2006	10.3984	0.4012	0.4349	14.6904	0.7423	0.1931
GridDehazeNet [42]	ICCV'19	22.7200	0.9071	0.0512	23.6556	0.8898	0.0681	19.2209	0.8061	0.1463	21.8925	0.8543	0.0951
FFA-Net [40]	AAAI'20	28.0407	0.9397	0.0398	28.1444	0.9160	0.0542	24.9337	0.8806	0.1080	26.4326	0.8960	0.0794
FCTF-Net [46]	GRSL'21	24.3520	0.9163	0.0594	24.2899	0.8932	0.0808	21.7059	0.8426	0.1522	22.6950	0.8634	0.1069
FSDGN [30]	ECCV'22	27.5011	0.9320	0.0383	27.1478	0.9134	0.0510	24.3580	0.8661	0.1046	25.9139	0.8838	0.0742
EMPF-Net [44]	TGRS'23	21.3244	0.8919	0.0812	24.1802	0.8931	0.0829	21.4463	0.8524	0.1454	22.5148	0.8624	0.0973
PSMB-Net [47]	TGRS'23	27.3374	0.9333	0.0422	27.0159	0.9094	0.0574	24.3324	0.8726	0.1105	25.7096	0.8891	0.0809
DehazeFormer [23]	TIP'23	27.3145	0.9309	0.0452	27.6252	0.9126	0.0568	24.1227	0.8625	0.1290	25.7962	0.8839	0.0873
ASTA [45]	GRSL'24	26.2978	0.9246	0.0473	26.1637	0.9064	0.0652	23.1796	0.8516	0.1389	24.7952	0.8733	0.0940
DEA-Net [43]	TIP'24	27.2645	0.9308	0.0389	27.5489	0.9099	0.0448	24.2043	0.8665	0.0886	25.7875	0.8843	0.0665
4KDehazing [41]	CVPR'21	27.4197	0.9305	0.0473	27.6984	0.9050	0.0645	24.3663	0.8622	0.1278	25.7836	0.8829	0.0886
Trinity-Net [3]	TGRS'23	25.1596	0.9014	0.0512	26.0094	0.9089	0.0632	22.9203	0.8491	0.1795	23.9387	0.8373	0.0742
Refusion [14]	CVPR'23	24.9612	0.8964	0.0496	26.1264	0.9132	0.0604	23.2159	0.8146	0.1352	23.8737	0.8324	0.0677
RDDM [12]	CVPR'24	25.1103	0.9036	0.0569	25.8964	0.9123	0.0502	23.8694	0.8264	0.1129	24.9379	0.9264	0.0671
RSHazeDiff [16]	TITS'24	26.4394	0.9004	0.0559	26.1943	0.8982	0.0594	24.0146	0.8640	0.1204	25.1140	0.8715	0.0645
DFG-DDM-Lite	–	28.1235	0.9407	0.0306	28.3619	0.9295	0.0461	25.3641	0.8815	0.0834	27.1058	0.8962	0.0562
DFG-DDM	–	28.4751	0.9498	0.0243	28.6189	0.9371	0.0397	25.9261	0.8861	0.0794	27.4179	0.8971	0.0518

TABLE III
AN ABLATION STUDY ON THE FREQUENCY MODULE OF DFG-DDM. THE BEST SCORE IS IN **BOLD**.

Module		SateHaze1k-all		
$\mathcal{FG}$	$\mathcal{FH}$	PSNR ↑	SSIM ↑	LPIPS↓
–	–	26.7049	0.8814	0.0697
✓	–	27.2068	0.8902	0.0399
–	✓	26.8193	0.8886	0.0496
✓	✓	**27.4179**	**0.8971**	**0.0379**

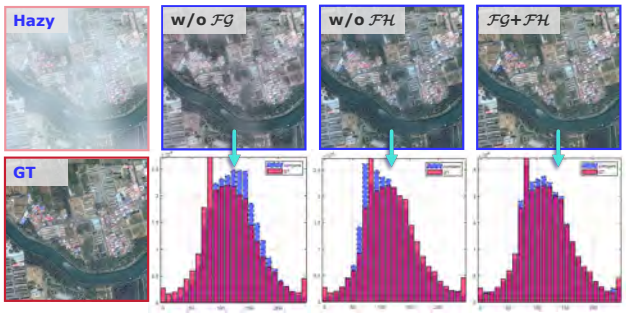

Fig. 8. Ablation study of the frequency module. The combined module ($\mathcal{FG}$ +$\mathcal{FH}$) results in images with better clarity and detail retention, as reflected in the histogram similarity to the ground truth. The greater the overlap of the histograms, the better the image recovery performance.

RSHaze, and SateHaze1k-all datasets, our method achieves improvements in PSNR of 0.159/1.239/0.985 over the best previous methods. In terms of LPIPS, our method outperformed a 44.9%/16.4%/22.1% percent improvement in error reduction. In the comparison of methods with similar parameters, our original model has a slightly higher parameter count than RSHazeDiff [16] but lower than Refusion [14]. Our lightweight model has a parameter of 33.7M, which is lower than other diffusion-based methods (*e.g.*, RDDM [12], Refusion [14] RSHazeDiff [16]). Nevertheless, our DFG-DDM-Lite consistently outperforms other methods in terms of results. These quantitative experimental results demonstrate the superior capability of our method in image restoration quality.

E. Ablation Study

We conduct ablation studies of two critical components of our method on both the frequency module and the loss function module. These experiments aim to objectively quantify individual contributions and validate the framework design rationale through controlled variable analysis.

1) Frequency Guidance and High-Frequency Enhancement: In this part, we first verify the importance of frequency domain guidance ($\mathcal{FG}$) for restoring image details, followed by assessing the effectiveness of high-frequency enhancement ($\mathcal{FH}$). As shown in Table III, the $\mathcal{FH}$ module alone offers minimal improvement, with optimal results achieved when combined with the $\mathcal{FG}$ module to leverage their complementary advantages. Fig. 8 illustrates the differences between various ablation modules and the ground truth. The original module shows noticeable blurriness and weak edges. $\mathcal{FG}$ improves structure and edge clarity, while $\mathcal{FH}$ sharpens edges and enhances texture, closely matching the ground truth.

2) Joint Loss Function: The second part of our ablation studies focuses on the design of the joint loss function, including pixel loss $\mathcal{L}_{\mathcal{P}}$, frequency loss $\mathcal{L}_{\mathcal{F}}$, and Charbonier loss $\mathcal{L}_{\mathcal{C}}$. Table IV presents the quantitative results. The joint loss function in DFG-DDM improves image recovery by approximately 1.5 dB on RSID and RSHaze datasets and nearly 2.5 dB on RHDRS dataset. This finding highlights the importance of the joint loss function. Furthermore, it emphasizes the wealth of frequency domain information inherent in the RHDRS dataset, underscoring the pivotal role of efficient recovery of frequency domain information in image dehazing.

Pixel Loss Only. Training solely with pixel loss leads to suboptimal restoration, causing weak edge definition, noise, and spectral distortion, especially around fine edges. While pixel-wise optimization effectively minimizes intensity discrepancies, it fails to preserve phase coherence in wavelet

TABLE IV
AN ABLATION STUDY FOR LOSS FUNCTION OF DFG-DDM. THE BEST SCORE IS IN **BOLD**.

Joint Loss Function			RHDRS			RSID			RSHaze		
$\mathcal{L}_{\mathcal{P}}$	$\mathcal{L}_{\mathcal{F}}$	$\mathcal{L}_{\mathcal{C}}$	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
✓	–	–	25.2532	0.8876	0.0768	26.5162	0.8889	0.0591	37.2524	0.8767	0.0704
✓	✓	–	26.3160	0.8993	0.0602	27.2101	0.9290	0.0394	37.9165	0.9065	0.0564
✓	–	✓	26.4186	0.8835	0.0597	26.9569	0.9045	0.0265	38.2483	0.9279	0.0321
✓	✓	✓	**27.8044**	**0.9082**	**0.0463**	**27.8517**	**0.9590**	**0.0103**	**38.8261**	**0.9549**	**0.0230**

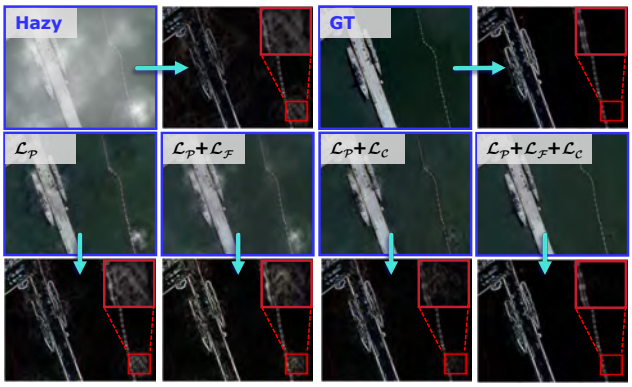

Fig. 9. Ablation study of the joint loss function. Visualization is displayed through canny edges, demonstrating the effectiveness of the joint loss function in detail recovery.

domains or edge-aware regularization. Moreover, the absence of frequency and Charbonier loss components hinders the ability to capture fine-grained structures.

Joint Loss Function. The joint loss function synergistically integrates multiple loss types, providing well-defined textures, color consistency, and effective noise suppression. Given the reliance of the diffusion process on frequency information, the frequency loss directly optimizes for differences in the frequency domain, aligning with the guiding mechanism of the diffusion process. This configuration of the joint loss function strengthens the network's generation direction, mitigating the potential loss of frequency-domain information that may occur when relying solely on pixel loss. Consequently, these improvements are evidenced by the sharper edges and richer details observed in the Canny edge visualizations, as depicted in Fig. 9.

V. CONCLUSION

In this article, we propose a diffusion generative model steered by frequency domain information to enhance the dehazing effect of remote sensing (RS) images. By integrating a deep Fourier transform into the noise prediction process, we enrich image generation with frequency domain control. Furthermore, we craft a joint loss function that simultaneously supervises losses in both the frequency and spatial domains. Extensive experiments prove that our proposed DFG-DDM achieves commendable results across multiple datasets. Additionally, we construct the RHDRS dataset to simulate more realistic and random haze distributions, thus expanding the pool of labeled data for RS image dehazing. In future research, we intend to design faster sampling mechanisms and integrate more efficiently with deep frequency domain information extraction to achieve higher quality and faster image dehazing results.

REFERENCES

[1] W. Han *et al.*, "A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 87–113, Aug. 2023.
[2] B. Huang, L. Zhi, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using sar image prior based on conditional generative adversarial networks," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1806–1813.
[3] K. Chi, Y. Yuan, and Q. Wang, "Trinity-net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, Jun. 2023.
[4] B. Li *et al.*, "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Aug. 2018.
[5] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Sep. 2010.
[6] E. J. McCartney, "Optics of the atmosphere: scattering by molecules and particles," *New York*, 1976.
[7] D. Berman, T. Treibitz, and S. Avidan, "Single image dehazing using haze-lines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 720–734, Nov. 2018.
[8] K. Chi, J. Li, W. Jing, Q. Li, and Q. Wang, "Neural implicit fourier transform for remote sensing shadow removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–10, Jun. 2024.
[9] C. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3d position embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5812–5820.
[10] B. Jiang *et al.*, "A dehazing method for remote sensing image under nonuniform hazy weather based on deep learning network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, Mar. 2023.
[11] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 23 593–23 606, Dec. 2022.
[12] J. Liu, Q. Wang, H. Fan, Y. Wang, Y. Tang, and L. Qu, "Residual denoising diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2773–2783.
[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840–6851, Dec. 2020.
[14] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1680–1691.
[15] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10 346–10 357, Jan. 2023.
[16] J. Xiong, X. Yan, Y. Wang, W. Zhao, X.-P. Zhang, and M. Wei, "Rshazediff: A unified fourier-aware diffusion model for remote sensing image dehazing," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 1, pp. 1055–1070, Nov. 2024.
[17] F. Luo, J. Xiang, J. Zhang, X. Han, and W. Yang, "Image super-resolution via latent diffusion: A sampling-space mixture of experts and frequency-augmented decoder approach," *arXiv:2310.12004*, 2023.
[18] K. Chi, S. Guo, J. Chu, Q. Li, and Q. Wang, "Rsmamba: Biologically plausible retinex-based mamba for remote sensing shadow removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–10, Jan. 2025.

[19] Z. Zhu, M. Xia, B. Xu, Q. Li, and Z. Huang, "Gtea: Guided taylor expansion approximation network for optical flow estimation," *IEEE Sens. J.*, vol. 24, no. 4, pp. 5053–5061, Jan. 2024.
[20] Z. Huang *et al.*, "T2EA: Target-aware taylor expansion approximation network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 5, pp. 4831–4845, Jan. 2025.
[21] D. Berman, S. Avidan *et al.*, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1674–1682.
[22] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4770–4778.
[23] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Trans. Image Process.*, vol. 32, pp. 1927–1941, Mar. 2023.
[24] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, pp. 980–993, Dec. 2021.
[25] Z. Huang *et al.*, "Joint analysis and weighted synthesis sparsity priors for simultaneous denoising and destriping optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 6958–6982, Mar. 2020.
[26] Z. Huang, Z. Zhu, Z. Wang, Y. Shi, H. Fang, and Y. Zhang, "Dgdnet: Deep gradient descent network for remotely sensed image denoising," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, Feb. 2023.
[27] Z. Huang *et al.*, "Rcst: Residual context sharing transformer cascade to approximate taylor expansion for remote sensing image denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, Jan. 2025.
[28] M. Cai, H. Zhang, H. Huang, Q. Geng, Y. Li, and G. Huang, "Frequency domain image translation: More photo-realistic, better identity-preserving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13 930–13 940.
[29] M. Fu, H. Liu, Y. Yu, J. Chen, and K. Wang, "Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 203–212.
[30] H. Yu, N. Zheng, M. Zhou, J. Huang, Z. Xiao, and F. Zhao, "Frequency and spatial dual guidance for image dehazing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 181–198.
[31] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, pp. 8780–8794, Dec. 2021.
[32] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16 293–16 303.
[33] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 624–632.
[34] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Dec. 2017.
[35] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, Jan. 2017.
[36] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Apr. 2017.
[37] Q. Guo, H.-M. Hu, and B. Li, "Haze and thin cloud removal using elliptical boundary prior for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9124–9137, Jul. 2019.
[38] J. Liu, R. W. Liu, J. Sun, and T. Zeng, "Rank-one prior: Real-time scene recovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8845–8860, Dec. 2022.
[39] J. Han, S. Zhang, N. Fan, and Z. Ye, "Local patchwise minimal and maximal values prior for single optical remote sensing image dehazing," *Inf. Sci.*, vol. 606, pp. 173–193, Aug. 2022.
[40] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, Feb. 2020, pp. 11 908–11 915.
[41] B. Xiao, Z. Zheng, Y. Zhuang, C. Lyu, and X. Jia, "Single uhd image dehazing via interpretable pyramid network," *Signal Processing*, vol. 214, p. 109225, Jan. 2024.
[42] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7314–7323.
[43] Z. Chen, Z. He, and Z.-M. Lu, "Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention," *IEEE Trans. Image Process.*, vol. 33, pp. 1002–1015, Jan. 2024.
[44] Y. Wen, T. Gao, J. Zhang, Z. Li, and T. Chen, "Encoder-free multi-axis physics-aware fusion network for remote sensing image dehazing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, Oct. 2023.
[45] Z. Cai, J. Ning, Z. Ding, and B. Duo, "Additional self-attention transformer with adapter for thick haze removal," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, Feb. 2024.
[46] Y. Li and X. Chen, "A coarse-to-fine two-stage attentive network for haze removal of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 10, pp. 1751–1755, Jul. 2020.
[47] H. Sun *et al.*, "Partial siamese with multiscale bi-codec networks for remote sensing image haze removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, Oct. 2023.
[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 586–595.

Junjie Li received the B.E. degree in software engineering from Zhengzhou University, Zhengzhou, China, in 2024. He is currently pursuing the M.S. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition, and remote sensing.

Kaichen Chi received the B.E. degree in electronic and information engineering and the M.E. degree in communication and information system from Liaoning Technical University, Huludao, China, in 2019 and 2022 respectively. He is currently working toward the Ph.D. degree in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include image processing and deep learning.

Yue Chang received the B.E. degree in earth environmental science and the Ph.D degree in earth and built environment science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2015 and 2022, respectively. He is currently a Postdoctoral Fellow with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing, deep learning, urban climatology.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/).