Spatial-Frequency Adaptive Remote Sensing Image Dehazing With Mixture of Experts

Hao Shen, Henghui Ding, *Member, IEEE*, Yulun Zhang, *Member, IEEE*, Xiaofeng Cong, *Student Member, IEEE*, Zhong-Qiu Zhao, *Member, IEEE*, and Xudong Jiang, *Fellow, IEEE*

***Abstract*— The feature modulation mechanism has been demonstrated to be particularly well-suited for efficient network design and is rarely explored in remote sensing dehazing tasks. Moreover, we observe distinct patterns in haze distribution across the low-frequency (LF) and high-frequency (HF) components of haze images from various datasets. However, existing research rarely investigated the potential solution in the frequency domain. In response, we propose a novel spatial-frequency adaptive network (SFAN), which is mainly built by the proposed mixture of modulation experts (MoME) and decoupled frequency learning block (DFLB). Different from the fixed feature modulation design used in other tasks, the MoME adopts the mixture-of-expert mechanism to dynamically learn diverse contextual features of various granularities and scales in a sample-adaptive manner and then utilize them to perform elementwise local feature modulation. This pure convolution architecture enables our network to have superior performance and efficiency tradeoffs. Furthermore, the DFLB is devised to facilitate the LF global haze removal and reconstruction of HF local texture information. At the micro level, we first utilize a mask extractor (ME) to generate the frequency mask from the input hazy image, then employ a dual-branch decoupled learning unit to boost frequency learning, and finally develop a mixture of fusion experts (MoFE) to achieve HF and LF feature interaction. Extensive experiments on publicly available dehazing datasets demonstrate that our network performs superior performance while incurring lower computational costs. Compared to the state-of-the-art approach (DEA-Net), SFAN achieves, an average, 0.83-dB PSNR improvement on five remote sensing datasets but consumes only 51% of the FLOPs. The code will be available at https://github.com/it-hao/SFAN.**

***Index Terms*— Decoupled frequency learning, image dehazing, mixture of modulation experts (MoME).**

Received 10 July 2024; revised 3 August 2024 and 26 August 2024; accepted 6 September 2024. Date of publication 12 September 2024; date of current version 7 October 2024. This work was supported in part by Anhui Special Support Plan for High-Level Talents and the Guangxi Key Research and Development Program under Grant AB22035022 and in part by China Scholarship Council under Grant 202306690021. *(Corresponding author: Zhong-Qiu Zhao.)*

Hao Shen and Zhong-Qiu Zhao are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: haoshenhs@gmail.com; z.zhao@hfut.edu.cn).

Henghui Ding is with the Institute of Big Data, Fudan University, Shanghai 200433, China (e-mail: henghui.ding@gmail.com).

Yulun Zhang is with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yulun100@gmail.com).

Xiaofeng Cong is with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: cxf_svip@163.com).

Xudong Jiang is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: EXDjiang@ntu.edu.sg).

Digital Object Identifier 10.1109/TGRS.2024.3458986

I. Introduction

IMAGES captured by remote sensing satellites often suffer from absorption and scattering effects caused by haze and thin clouds, ultimately leading to image degradation. The low quality of these images cannot directly be used for subsequent high-level computer vision tasks, such as object detection [1], semantic segmentation [2], image captioning [3], and video grounding [4], [5]. Therefore, developing an effective method for removing haze and thin clouds from single remote sensing images is crucial. Nonetheless, dehazing based on a single image remains challenging due to the ill-posed property.

Over the past few years, many researchers have devoted themselves to this task and proposed many algorithms, mainly including prior-based and data-driven methods. The former is primarily based on the atmospheric scattering model (ASM), incorporating diverse prior assumptions, such as dark channel prior (DCP) [6], haze lines [7], and dark-object subtraction [8], into the network to learn jointly. However, prior-based techniques often yield subpar dehazing outcomes when confronted with dense and nonhomogeneous hazy environments. This is due to the intricate nature of accurately estimating multiple haze parameters using prior-based methods.

The data-driven methods leverage substantial amounts of training data and powerful representation capabilities of neural networks to train models, thereby dominating remote sensing image dehazing tasks. Earlier deep networks [13], [14], [15] first leverage powerful feature representations to estimate ambient light and medium transmission and then utilize the ASM [16] to estimate hazy-free images. Recently, more methods [10], [11], [17], [18], [19] tend to adopt an end-to-end manner to produce hazy-free images. However, with the excellent characterization capability of the vision Transformer (ViT), more and more haze removal algorithms [11], [20], [21], [22], [23] incorporate self-attention as the fundamental building block to design the network. For example, DeHamer [20] utilized the Transformer features to modulate the convolutional features, achieving feature consistency. DehazeFormer [22] redesigned a critical structure of Swin Transformer [24] and introduced the spatial information aggregation mechanism for better both natural and remote sensing image dehazing. MBTFormer [21] utilized the Taylor expansion and multi-scale design to build a novel linearized Transformer network. TrinityNet [11] wisely incorporated prior information into the convolutional neural network (CNN) and Swin Transformer for better estimating haze parameters. Despite the significant

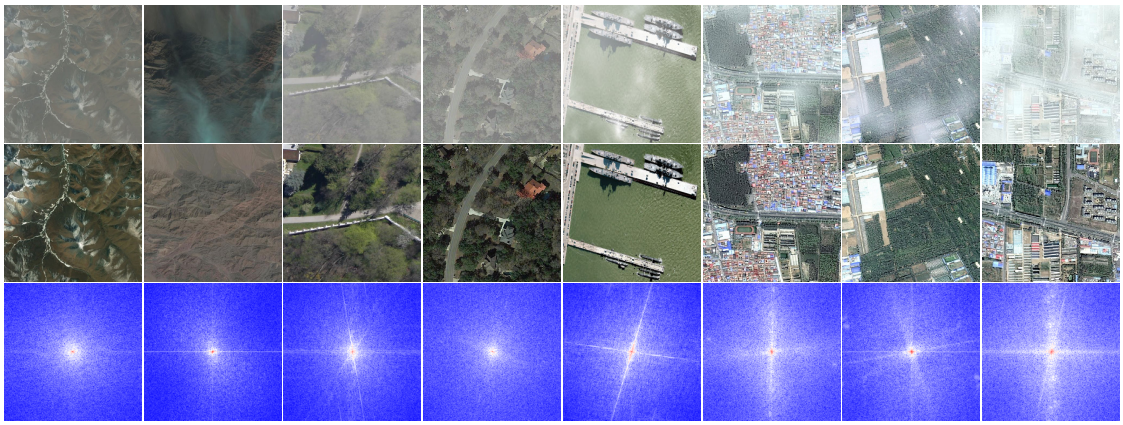

Fig. 1. From the top to bottom represents the hazy images (top), clear images (middle), and the Fourier spectra (bottom) of residual images obtained by subtracting the hazy images from the clear images for different remote sensing datasets. For different datasets, haze affects LF and HF information in images differently. Therefore, it is necessary to conduct adaptive learning for frequency features.

advancements achieved by these methodologies, an additional challenge emerges from the stringent efficiency requirements.

More recently, modulation mechanisms [25], [26] show that pure convolutional networks could also achieve superior performance and are computationally efficient compared to Transformers. These methods usually employ a uniform modulation framework across all layers. However, the scale and structural information of the features change as the network depth varies, and utilizing a single paradigm to modulate features at each layer restricts the diversity of learned features. Especially for remote sensing images from large-area Earth observation scenarios and large-range imaging, there exists scale diversity and channel redundancy of features [27]. Therefore, using a single, monolithic model to deal with the dehazing task is challenging. The mixture-of-expert (MoE) [28] mechanism leverages expert knowledge of various components to learn different aspects and adopts a dynamic network to integrate them, enabling sample-adaptive targets. By introducing the MoE mechanism into the feature modulation scheme, we can achieve adaptive feature modulation and enhance the network's generalization capabilities.

In addition, recent studies [29], [30], [31], [32], [33] have demonstrated the effectiveness of customized learning for frequency features. FMSR [30] devised an adaptive frequency-assisted Mamba for remote sensing image super-resolution to grasp frequency feature dependencies. FMRNet [32] designed a frequency mutual revision deraining network to eliminate rain perturbation while preserving background textures in frequency space. SFNet [33] combined dynamic convolution with multibranch architecture to perform decoupled learning of high-frequency (HF) and low-frequency (LF) information for image restoration. Inspired by these works, we also count the distribution of haze information in the frequency domain on different remote sensing dehazing datasets. As shown in Fig. 1, we observe that the haze distribution in the low- and high-frequency components of haze images from different datasets is different. For instance, in the RICE [9], DHID, and LHID [10] datasets, haze degradation primarily affects low-frequency information of images, whereas, in the RSID [11] and StateHaze1K [12] datasets, both low- and high-frequency information is contaminated by haze. Therefore, it is necessary to conduct in-depth research on the frequency characteristics and further deal with HF and LF features in a targeted manner to reduce learning difficulty and achieve effective decoupled learning.

We attempt to tackle the remote sensing image dehazing task by incorporating the spatial and frequency domains and propose a spatial-frequency adaptive network (SFAN). To adaptively model the contextual relationships while removing haze from the hazy images, we delicately develop two core designs: modulation expert block (MEB) in the spatial domain and decoupled frequency learning block (DFLB) in the spatial-frequency domain. Specifically, we devise a mixture of modulation experts (MoME) as the fundamental component within the MEB to adaptively select the most suitable cross-spatial or cross-channel modulation experts and explicitly achieve contextual feature modeling. Each modulation expert employs elementwise operations to enhance high-dimensional feature interaction, thereby improving information transformation. The DFLB is composed of a mask extractor (ME), a decoupled learning unit, and a mixture of fusion experts (MoFE). First, the ME is responsible for generating frequency masks that partition the image into high-frequency and LF parts. Second, the decoupled learning unit adopts a dual-branch structure to facilitate the independent learning of low- and high-frequency features with deep hierarchical features using cross-attention (CA), enabling effective haze removal and reconstruction of delicate texture details. Finally, the MoFE is responsible for the adaptive interaction of HF and LF features. With the aid of the proposed core components, our method can effectively promote haze removal and present better generalization. Our contributions can be summarized as follows.

1) We devise an SFAN from the perspective of the spatial-frequency domain to tackle the remote sensing image dehazing task. Comprehensive experiments verify that our SFAN achieves superior performance on ten public dehazing datasets while being highly efficient.
2) An MEB is customized to adaptively aggregate contextual information, which adopts a dynamic structure to

extract diverse features. This is a significant attempt to introduce MoE in the remote sensing image dehazing task.

3) A DFLB is developed to facilitate haze removal and enhance high-frequency information reconstruction.

II. Related Works

A. Prior-Based Image Dehazing

Such methods typically use manual priors to remove haze from empirical observations. As for the NSI dehazing task, He et al. [6] first proposed the DCP with the haze imaging model to directly estimate haze-free images. Immediately afterward, Zhu et al. [34] utilized a linear model to build the scene depth of haze images based on color attenuation priors, further achieving the haze removal by the obtained depth information. Berman et al. [35] proposed a nonlocal dehazing method since the pixels observed in a given cluster are distributed across the entire image and located at different distances from the camera. For RSI tasks, Li et al. [36] proposed a homomorphic filtering and sphere model to improve DCP based on the observation that haze is mainly located in the LF component of an image. Shen et al. [37] developed a globally nonuniform atmospheric light model and introduced a bright pixel index to represent spatially varied atmospheric light and to extract local bright surfaces. However, these prior-based methods may lead to transmission estimation errors due to imprecise prior information, further producing dehazing images with slight color distortion or other degradation phenomena.

B. Deep Learning-Based Image Dehazing

Different from prior-based methods, early deep learning-based methods [13], [14], [15], [38] often utilize CNNs to estimate the transmission map and atmospheric light, respectively, and then use the ASM [16] to obtain haze-free images. Nevertheless, estimating the atmospheric light value and transmission map can never fully reflect the complex atmospheric conditions found in nature. As a result, the dehazed images often suffer from degradation issues.

Recently, designing an end-to-end dehazing network to predict haze-free images has become mainstream since it can achieve superior performance. For NSI dehazing, Ren et al. [39] gated and fused multiple enhanced images derived from initial haze images to generate haze-free images and introduced the multiscale approach to train the network more accurately. Qin et al. [40] designed a very deep network based on feature attention and residual learning. Dong et al. [41] proposed the MSBDN, which employs a dense feature fusion module based on the back projection scheme and boosting strategy to build the network. Wu et al. [17] utilized contrastive learning to ensure that the restored image is closer to the clear image and farther away from the hazy image in the representation space. Zhang et al. [42] utilized a mutual promotion framework that treats depth estimation and image dehazing independently but optimizes them by a dual-task interaction mechanism. For RSI dehazing, Li and Chen [19] devised a two-stage dehazing network and adopted a coarse-to-fine strategy to train the network. Zhang and Wang [10] presented a collaborative criterion and a shared-weight Siamese network structure to improve the robustness and generalization performance for remote-sensing dehazing. Sun et al. [43] designed a Siamese network to improve the constraint ability of the hazy area and utilized the multiscale information to improve the reconstruction ability of the network for the color and texture. Wen et al. [18] introduced an innovative encoder-free design to accomplish the lightweight and efficient target. Du et al. [44] incorporated the ASM with CNN architecture to realize the joint optimization of physical parameters and model parameters. Lihe et al. [45] developed a physics-aware network to achieve effective haze removal for remote sensing images, providing a method with physical interpretability.

More recently, due to the powerful global modeling capabilities of Transformer, most of the prevailing dehazing methods based on this architecture have been proposed. Among them, Guo et al. [20] first combined the local representation capability of CNN and the global context modeling capability of Transformer to improve performance significantly. Chi et al. [11] integrated the prior information into the Swin Transformer [24] to rich perceptual details, achieving accurate remote sensing image dehazing. Song et al. [22] modified the normalization layer in Swin Transformer [24] and introduced spatial information aggregation schema to improve the capability of multihead self-attention, achieving superior dehazing performance in both NSI and RSI tasks. However, these methods consume much computation when calculating self-attention, which cannot satisfy the demand for efficiency. Yu et al. [46] utilized the Fourier transform to decouple the amplitude and phase information and embedded them into CNN, realizing haze removal and structure construction efficiently. Shen et al. [31] utilized mutual information constraints to accomplish complementary learning from spatial and frequency domains.

C. Mixture-of-Experts

MoE can be regarded as a divide-and-conquer technique, where the problem space is partitioned among several submodels called experts. Consequently, the output of the MoE is a weighted blend of the outputs from various experts, achieving dynamic adjustment. The sparse MoE generally employs a routing mechanism to dynamically and adaptively direct input to a subset of these experts, enhancing model capacity without increasing computational complexity, which became popular in natural language processing (NLP). It usually appears as a fundamental model component, such as the FFN layer in Transformer architecture [47], [48].

In computer vision, He et al. [49] introduced the sparse MoE into the pan-sharpening task to boost the decoupling learning of HF and LF features. Cao et al. [50] proposed a multimodal gated mixture of local-to-global experts to realize reliable infrared and visible image fusion. Yang et al. [51] leveraged the prior from the vision-language model to choose suitable experts for restoring degraded images. In this article, we propose a remote sensing image dehazing method based on the MoE mechanism for the first time.

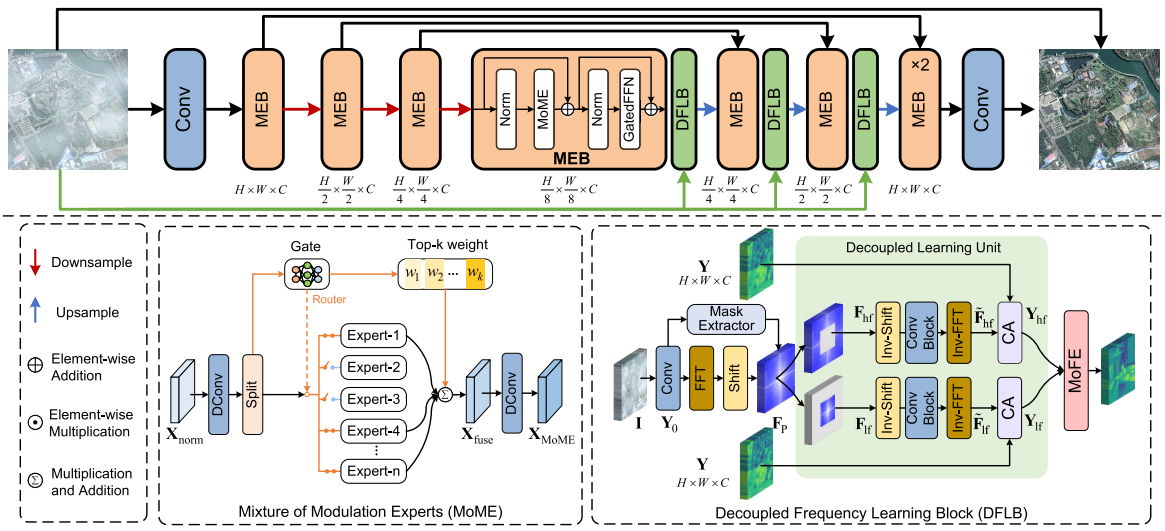

Fig. 2. Overall framework and its components. Specifically, the proposed SFAN adopts an asymmetric encoder–decoder design, where the MEB and DFLB are the core components.

III. Methods

A. Overall Architecture

The overall pipeline of the SFAN is presented at the top of Fig. 2. Given a hazy image $\mathbf{I}_{\text{hazy}} \in \mathbb{R}^{H\times W\times 3}$, where the $H \times W$ refers to the spatial size, the network initially employs a 3×3 convolution to extract shallow hazy features. Subsequently, a hierarchical encoder–decoder design is used to extract cross-scale features. In the encoder, each layer is constructed based on the proposed MEB. Meanwhile, the spatial dimension of the feature maps is gradually reduced by two times, while the channel dimension is progressively increased by two times. Therefore, the lowest scale latent feature representation $\mathbf{Y}_{\text{low}}$ can be obtained. The decoder starts with the low-resolution feature $\mathbf{Y}_{\text{low}}$ and progressively recovers the high-resolution hazy-free features. Thus, the spatial dimension gradually increases while the channel dimension gradually decreases. Different from the encoder, we insert a DFLB between every two levels of the decoder to assist the feature learning. To switch the feature scale, we employ the pixel-unshuffle and pixel-shuffle operations combined with convolution to implement the feature downsampling and upsampling. Similar to previous work [21], [31], [52], skip connection is also used to link the features from the corresponding layer of the encoder and decoder. Finally, a 3×3 convolution is deployed in high-resolution deep features $\mathbf{Y}_{\text{deep}}$ to generate a residual image to which the hazy image is added to obtain the hazy-free clean image $\mathbf{I}_{\text{out}}$. Next, we will describe two key components: MEB and DFLB.

B. Modulation Expert Block

The self-attention mechanism in Transformers usually suffers from quadratic complexity over the number of visual tokens, making it inefficient in terms of both parameters and computation. To this end, we consider inheriting the Transformer-like architecture yet adopting a different manner to model feature dependencies efficiently. The MoME is designed to mine the contextual information of features with varied scales and channels using the MoE mechanism. Formally, given the input features $\mathbf{X}$, the process of the MEB can be defined as

$$\mathbf{X}_1 = \mathbf{X} + \text{MoME}(\text{LN}(\mathbf{X})) \tag{1}$$

$$\mathbf{X}_2 = \mathbf{X}_1 + \text{GatedFFN}(\text{LN}(\mathbf{X}_1)) \tag{2}$$

where $\text{LN}(\cdot)$ denotes the layer normalization operation, GatedFFN [53] is used for refining contextual information, and $\mathbf{X}_1$ and $\mathbf{X}_2$ are the outputs of the MoME and GatedFFN modules, respectively.

1) Mixture of Modulation Expert: The MoME contains a gate network and N specially configured modulation experts $\{\mathcal{E}_i, \mathcal{E}_2, \ldots, \mathcal{E}_N\}$. The gate produces different weights for feature modulation, so the most suitable experts are chosen based on the input maps. As observed, a feature projection layer at the head and tail of the module is employed in the block. Specifically, a 3×3 depthwise convolution (DConv) followed by the channel split operation is performed on the normalized features $\mathbf{X}_{\text{norm}}$ to realize feature projection and division, denoting the resulting features as $\mathbf{X}_{\text{q}}$ and $\mathbf{X}_{\text{v}}$. Then, $\mathbf{X}_{\text{q}}$ is inputted into the gate network to generate the weight, which can be denoted as

$$\mathbf{X}_{\text{p}} = \text{GAP}(\mathbf{X}_{\text{q}}) + \text{GMP}(\mathbf{X}_{\text{q}}) \tag{3}$$

$$\mathbf{V} = \mathbf{X}_{\text{p}} \cdot \mathbf{A}_{\text{g}} + \mathcal{N}(0, 1) \cdot \text{SoftPlus}(\mathbf{X}_{\text{p}} \cdot \mathbf{A}_{\text{noise}}) \tag{4}$$

$$\mathbf{W} = \text{Softmax}(\text{TopK}(\mathbf{V})). \tag{5}$$

$\text{GAP}(\cdot)$ and $\text{GMP}(\cdot)$ are average pooling and maximum pooling operations, which are utilized to obtain local feature descriptors $\mathbf{X}_{\text{p}}$. $\mathbf{A}_{\text{g}}$ and $\mathbf{A}_{\text{noise}}$ denote learnable weight matrices, which are used together with $\mathbf{X}_{\text{p}}$ to generate $\mathbf{V}$. This series of operations aims to map the input into N router logits for expert selection. Finally, applying the TopK algorithm, the k positions with the highest weight are selected, and the remaining unselected expert weights are set to negative infinity. This way, the weights $\mathbf{W}$ of selected k experts can be obtained using the Softmax, and the unselected expert weights are assigned to zero. Here, utilizing learnable noise ensures

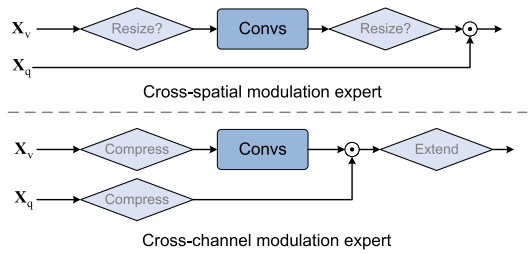

Fig. 3. Our customized cross-spatial and cross-channel modulation experts.

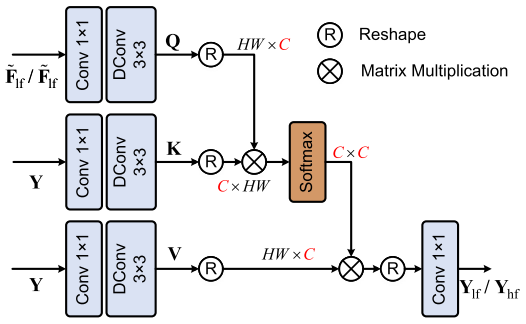

Fig. 4. Structure of the CA mechanism [55].

that the probability of selecting each expert is equal. The final output of the MoME is calculated as follows:

$$\mathbf{X}_{\text{fuse}} = \sum_{i=1}^{N} \mathbf{W}_i \cdot \mathcal{E}_i(\mathbf{X}_{\text{q}}, \mathbf{X}_{\text{v}}) \tag{6}$$

where $\mathbf{X}_{\text{fuse}}$ is the aggregated output and $\mathcal{E}_i(\cdot)$ denotes ith modulation expert function. Finally, the modulated features are passed through a convolution for interaction, producing the output denoted as $\mathbf{X}_{\text{MoME}}$. In Section III-B2, we will elaborate on the specifics of the various modulation experts.

2) Modulation Expert: To further delve into the intricacies of the interdependencies among the extracted features, we develop two forms of modulation experts, as shown in Fig. 3: one based on cross-spatial and the other cross-channel. Based on the diverse candidate experts, we can efficiently tap the maximum potential.

a) Cross-spatial modulation: To efficiently aggregate spatial-varied contextual features: 1) we directly perform feature extraction for $\mathbf{X}_{\text{v}}$ and 2) we first perform feature downsampling, then perform context aggregation on the low-resolution features, and finally apply the interpolation algorithm to upsample context features. The obtained coarse or fine-grained context can be denoted as $\mathbf{X}_{\text{c}}$ and then utilized to achieve elementwise modulation for query feature $\mathbf{X}_{\text{q}}$. Here, we adopt the same feature aggregation scheme as FocalNet [26].

b) Cross-channel modulation: As the network depth changes, the number of feature channels required for each layer differs. Therefore, we devise a cross-channel modulation schema to capture each layer's most informative feature representation and improve the efficiency. Precisely, we first compress the encoded features $\mathbf{X}_{\text{v}}$ along the channel dimension, yielding the low-rank encoded feature $\mathbf{X}_{\text{v}}^{\text{c}}$. Similarly, the low-rank query features $\mathbf{X}_{\text{q}}^{\text{c}}$ can be attained. Depending on the channel compression rate (R), various low-rank features can be collected. Next, we employ the same scheme as above to contextualize the encoded features and then leverage it to perform elementwise modulation for $\mathbf{X}_{\text{q}}^{\text{c}}$ in low-dimensional feature space. Finally, an additional convolution operation is applied to restore the features to their original dimensionality.

C. Decoupled Frequency Learning Block

As shown in Fig. 2, the DFLB contains two inputs: the hazy image $\mathbf{I}$ and the extracted hierarchical features $\mathbf{Y}$. To facilitate the decoupling of HF and LF information, the initial step involves determining the low-high-frequency boundary represented as a 2-D mask, and the second is to combine it with the transformed features using Fourier transform to obtain the corresponding frequency features.

1) Mask Extractor: To produce the mask with the same dimension as $\mathbf{Y}$, a 3×3 convolution is applied on the hazy image $\mathbf{I}$ to achieve feature alignment, resulting in $\mathbf{Y}_0$. As for the first step, we devise a lightweight ME similar to [54] to separate the spectra of the input hazy image. The ME is composed of a global average pooling, two 1×1 convolutions with the GELU activation, and a sigmoid function. The detailed operations can be denoted as

$$\alpha, \beta = \text{sigmoid}(\text{Conv}_2(\text{GELU}(\text{Conv}_1(\text{GAP}(\mathbf{F}_{\text{A}}))))). \tag{7}$$

Based on this, two scalars, α and β, can be obtained from zero to one. Subsequently, the LF mask is obtained by setting the value of the region specified in the following equation to ones and the remaining region to zeros:

$$\mathbf{M}_{\text{lf}}\left[\frac{H}{2} - \frac{\alpha \cdot H}{2} : \frac{H}{2} + \frac{\alpha \cdot H}{2}, \frac{W}{2} - \frac{\beta \cdot W}{2} : \frac{W}{2} + \frac{\beta \cdot W}{2}\right] = 1. \tag{8}$$

Similarly, the high-frequency mask $\mathbf{M}_{\text{hf}}$ can be attained by filling in zeros in the designated rectangle area and ones in the rest. As for the second step, the Fourier transform operation and Fourier shifting operation are sequentially performed on the feature maps $\mathbf{Y}_0$ to obtain the spectrum feature $\mathbf{F}_{\text{P}}$. Ultimately, by multiplying the corresponding mask with spectrum features, we can attain decoupled LF and high-frequency features in the Fourier domain, denoted as $\mathbf{F}_{\text{lf}}$ and $\mathbf{F}_{\text{hf}}$, respectively.

2) Decoupled Learning Unit: Inspired by recent studies [31], [46], which indicate that haze information of hazy images predominantly resides in the Fourier amplitude spectrum, while the phase spectrum conveys more structural information. Thus, we consider performing spectral learning (SL) for frequency features and proceed as follows: for the extracted HF and LF features, we apply a convolution block that consists of two 1×1 convolutions and a GELU layer to the amplitude of $\mathbf{F}_{\text{lf}}$ and the phase of $\mathbf{F}_{\text{hf}}$, respectively. Next, we conduct the inverse Fourier operation to convert them to the spatial domain, denoted as $\tilde{\mathbf{F}}_{\text{lf}}$ and $\tilde{\mathbf{F}}_{\text{hf}}$.

To better exploit the complementarity of decoupled frequency features and hierarchical spatial features $\mathbf{Y}$, we utilize

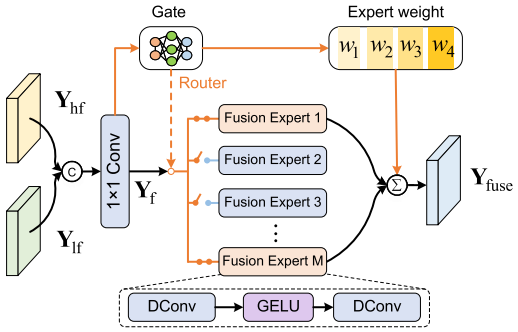

Fig. 5. Structure of MoFE.

the CA [55] mechanism, as shown in Fig. 4, which operates on channel dimension to achieve feature interaction. Taking LF features as an example, to be specific, we formulate the frequency prior $\tilde{\mathbf{F}}_{lf}$ as the *query* ($\mathbf{Q}$) features and the spatial feature $\mathbf{Y}$ as the *key* ($\mathbf{K}$), *value* ($\mathbf{V}$). All these transformed features are extracted using a 1×1 convolution followed by a 3×3 DConv, and then, they are reshaped to facilitate attention calculation. Therefore, the CA between the frequency and spatial features can be obtained by

$$\mathbf{Y}_{lf} = \text{Attention}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}}) = \text{Softmax}\left(\frac{\hat{\mathbf{Q}}\hat{\mathbf{K}}^{\mathsf{T}}}{\alpha}\right)\hat{\mathbf{V}} \tag{9}$$

where α is a learnable parameter that regulates the magnitude of the dot product. $\mathbf{Y}_{lf}$ represents the output features guided by the LF features $\tilde{\mathbf{F}}_{lf}$. Likewise, we can obtain the integrated features $\mathbf{Y}_{hf}$ under the guidance of the high-frequency features $\tilde{\mathbf{F}}_{hf}$.

3) Mixture of Fusion Experts: With the features from dual branches, as shown in Fig. 5, we propose an adaptive fusion schema based on the MoE mechanism, termed the MoFE, which can dynamically adjust its structure and parameters when integrating different frequency features. Like MoME, the MoFE comprises a gate network and multiple expert modules. The structure of each expert consists of two cascaded DConvs with a GELU activation. Given the low- and high-frequency features, we first concatenate them and subsequently perform a convolutional operation, resulting in the initial fused features $\mathbf{Y}_f$, which are then fed into the gate network and expert modules. The generation of weight $\mathbf{W}$ follows the same process as in Section III-B. The final output of the DFLB is a linearly weighted combination of each expert's output, modulated by the respective gating weights. The formalization is given as follows:

$$\mathbf{Y}_{fuse} = \sum_{j=1}^{M} \mathbf{W}_j \cdot \mathcal{F}_j(\mathbf{Y}_f) \tag{10}$$

where $\mathbf{Y}_{fuse}$ denotes the adaptive fused features and $\mathcal{F}_j(\cdot)$ represents the jth fusion expert.

D. Loss Function

Our SFAN utilizes a spatial-frequency hybrid architecture to construct the end-to-end network. Consequently, the loss function comprises both spatial and frequency loss components. Let $\mathbf{I}_{gt}$ and $\mathbf{I}_{out}$ denote the clear image and the dehazed image, respectively. The loss function can then be expressed as follows:

$$\mathcal{L}_{spa} = \left\|\mathbf{I}_{out} - \mathbf{I}_{gt}\right\|_1 \tag{11}$$

$$\mathcal{L}_{fre} = \left\|\mathcal{A}(\mathbf{I}_{out}) - \mathcal{A}(\mathbf{I}_{gt})\right\|_1 + \left\|\mathcal{P}(\mathbf{I}_{out}) - \mathcal{P}(\mathbf{I}_{gt})\right\|_1 \tag{12}$$

$$\mathcal{L}_{total} = \mathcal{L}_{spa} + \alpha\mathcal{L}_{fre} \tag{13}$$

where $\mathcal{L}_{spa}$ and $\mathcal{L}_{fre}$ represent spatial and frequency losses, respectively, $\mathcal{L}_{total}$ is the final loss, and $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ denote the Fourier amplitude and phase components, respectively. The tradeoff factor α is empirically set to 0.05.

IV. Experiments

A. Experimental Setup

1) Datasets:

1) *SateHaze1K [12]:* The dataset is categorized into three levels: SateHaze1K-thin, SateHaze1K-moderate, and SateHaze1K-thick. Each subdataset comprises 400 pairs of synthetic remote sensing hazy images, with 320 pairs allocated for training, 35 pairs for validation, and 45 pairs for performance evaluation.
2) *RSID :* This dataset was created by Chi et al. [11], which contains a total of 1000 paired images. We randomly selected 900 images to train and evaluated the model's performance using the remaining 100 images.
3) *Light Hazy Image Dataset and Dense Hazy Image Dataset [10]:* Based on the haze thickness, we have named the datasets the light hazy image dataset (LHID) and the dense hazy image dataset (DHID). The LHID comprises 14 490 paired images for training and 500 paired images for evaluation. The DHID includes 30 517 paired images for training and a separate set of 500 paired images for evaluation.
4) *RICE1 and RICE2:* The RICE [9] dataset was collected on Google Earth, thus covering different types of Earth's surface, such as urban scenes, deserts, mountains, and oceans. For RICE1, we randomly chose 402 images for training and 98 images for evaluation, and for RICE2, we used 590 images for training and 146 images for testing.
5) *Dense-Haze [60] and NH-HAZE [61]:* This former contains homogeneous haze and serves as the official dataset of the NTIRE 2019 challenge. The latter contains nonhomogeneous haze and serves as the official dataset of the NTIRE 2020 challenge. Both consist of 45 training image pairs, five validation image pairs, and five test image pairs and are used to evaluate natural image dehazing.
6) *RRSD300 [62]:* This dataset is composed of a total of 300 real-world remote sensing hazy images collected from the Microsoft Bing and DIOR datasets.

2) Implementation Details: At the macro level, the architecture of our SFAN employs a four-level encoder–decoder structure, with the same number of MEB at each level. In addition, we insert a DFLB between two sequential MEBs in the decoder, so there are three DFLB modules in total. The channel numbers from the first to the fourth level are 32,

TABLE I

QUANTITATIVE RESULTS OF REMOTE SENSING IMAGE DEHAZING ON DHID [10], LHID [10], RICE1 [9], RICE2 [9], AND RSID [11] DATASETS. THE FLOPS ARE CALCULATED ON THE 256 × 256 × 3 IMAGE PATCH. THE BEST RESULTS ARE IN BOLD

Methods	Venue	DHID [10]		LHID [10]		RICE1 [9]		RICE2 [9]		RSID [11]		Params.	FLOPs
		SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	(M)	(G)
SDCP [36]	GRSL'18	0.182	11.392	0.160	12.241	0.519	14.086	0.376	14.896	0.788	17.206	-	-
MinVP [56]	INS'18	0.178	11.370	0.196	14.381	0.819	18.889	0.544	16.505	0.803	17.891	-	-
IDeRs [57]	PROC'18	0.144	8.734	0.109	9.606	0.530	13.957	0.377	13.011	0.647	13.515	-	-
DCINet [10]	TGRS'22	0.938	28.865	0.941	28.320	0.958	29.812	0.793	25.628	0.928	24.499	26.51	26.86
EMPFNet [18]	TGRS'23	0.920	25.502	0.958	30.232	0.969	31.552	0.846	28.999	0.912	21.708	0.30	2.90
FCTFNet [19]	GRSL'20	0.907	24.929	0.964	32.047	0.975	31.407	0.872	31.526	0.919	22.976	0.16	10.05
PSMBNet [43]	TGRS'23	0.935	28.333	0.965	31.870	0.939	27.979	0.847	28.134	0.925	23.262	12.48	98.19
TrinityNet [11]	TGRS'23	0.913	26.189	0.938	27.219	0.959	29.659	0.856	28.836	0.927	24.196	20.14	30.78
4KDehazing [58]	CVPR'21	0.933	27.466	0.953	29.200	0.941	27.222	0.850	27.529	0.915	23.385	34.55	103.89
AECRNet [17]	CVPR'21	0.938	28.614	0.964	32.350	0.921	24.055	0.770	26.257	0.889	21.839	2.61	43.04
DeHamer [20]	CVPR'22	0.935	28.040	0.941	28.233	0.947	30.956	0.851	29.752	0.848	22.369	132.4	59.67
FSDGN [46]	ECCV'22	0.936	28.793	0.967	33.509	0.982	33.420	0.858	31.017	0.946	25.398	2.73	15.59
MITNet [31]	MM'23	0.938	27.668	0.952	29.502	0.979	33.450	0.873	31.086	0.935	24.243	2.73	16.42
DehazeFormer [22]	TIP'23	0.931	27.525	0.962	31.606	0.982	36.464	0.886	33.835	0.946	25.622	2.51	25.79
PhDnet-S [45]	INFFUS'24	0.931	26.898	0.967	33.091	0.982	36.246	0.895	33.933	0.950	25.894	5.27	17.06
DEA-Net [59]	TIP'24	0.933	27.318	0.967	33.451	0.983	35.244	0.891	34.010	0.951	25.918	3.65	32.23
SFAN (ours)	-	**0.942**	**29.173**	**0.970**	**34.029**	**0.985**	**36.653**	**0.897**	**34.095**	**0.952**	**26.135**	3.95	16.41

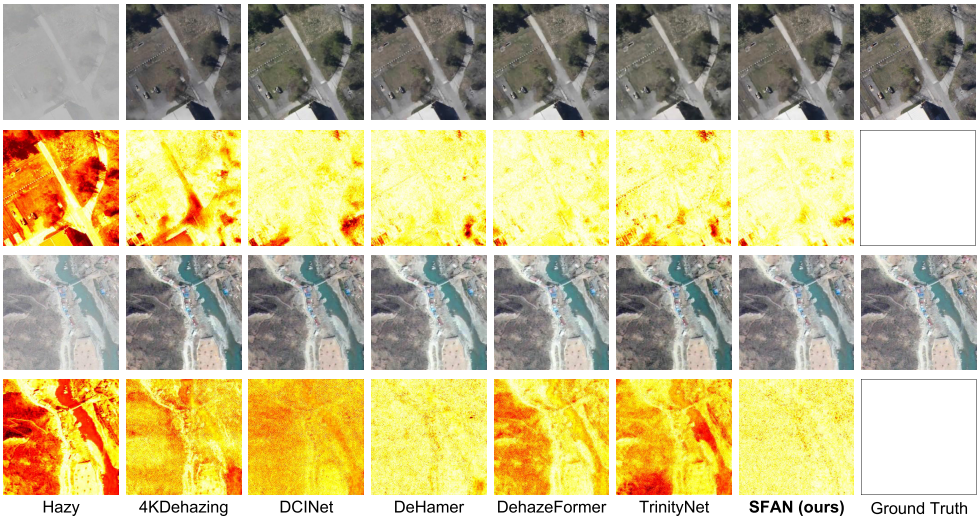

Fig. 6. Visual results and error maps on DHID and LHID datasets [10]. The error map is generated by computing the difference between the dehazed image produced by each method and the corresponding ground-truth image. It is clear that our method produces better dehazing results and smaller errors than competitors.

64, 128, and 256. At the micro level, we set six modulation experts for each MoME and sparsely select three experts for integration. For each MoFE, we set four fusion experts and sparsely select two experts for integration. For training, the ADAM optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used. For the Dense-Haze [60] and the NH-HAZE [61] datasets, the learning rate is initialized to 2×10^{-4} and gradually decreases to 1×10^{-6}. For other datasets, the learning rate is initialized to 2×10^{-4} and linearly decays by a factor of 0.9 every 10 epochs. We implement the proposed models on the PyTorch framework with a single NVIDIA 4090Ti GPU to train all models on all datasets. The batch and patch sizes are set to 4 and 256 × 256, respectively.

3) Evaluation Metrics: Two widely acknowledged metrics, peak signal noise ratio (PSNR) and structural similarity (SSIM), are utilized to evaluate all datasets to access our proposed methods. The code implementation is followed with FFA-Net [40] and MSBDN [41].

B. Synthetic Remote Sensing Image Dehazing

To demonstrate the efficacy of the proposed SFAN, we first conduct a comprehensive analysis with other RSI and NSI dehazing methods. The RSI dehazing methods consist of SDCP [36], MinVP [56], IDeRs [57], DCINet [10], EMPFNet [18], FCTFNet [19], PSMBNet [43], TrinityNet [11], and PhDnet-S [45]. The NSI dehazing methods include 4KDehazing [58], AECRNet [17], DeHamer [20], FSDGN [46], MITNet [31], DehazeFormer [22], and DEA-Net [59].

In Table I, we present the quantitative evaluation results on DHID [10], LHID [10], RICE1 [9], RICE2 [9], and RSID [11] datasets, from which we find that the proposed method

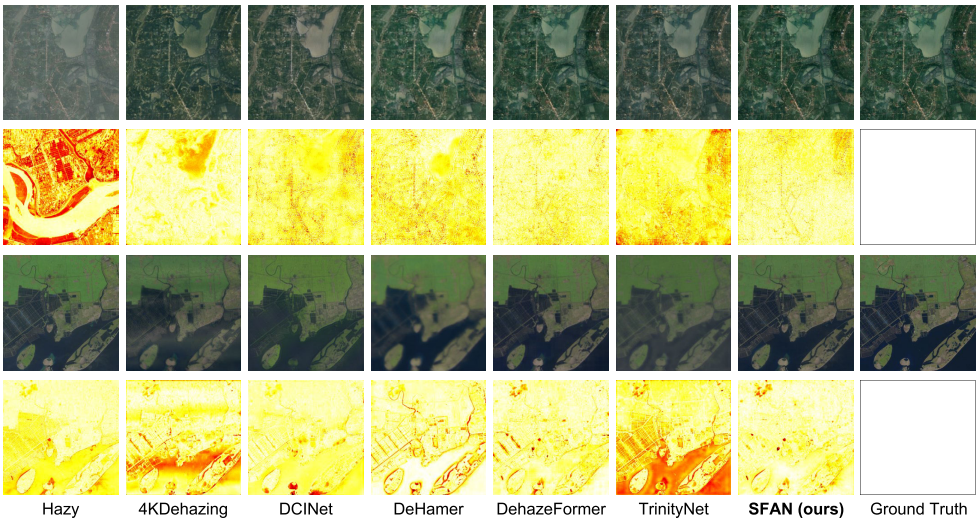

Fig. 7. Visual results and error maps on RICE1 and RICE2 datasets [9]. The error map is generated by computing the difference between the dehazed image produced by each method and the corresponding ground-truth image.

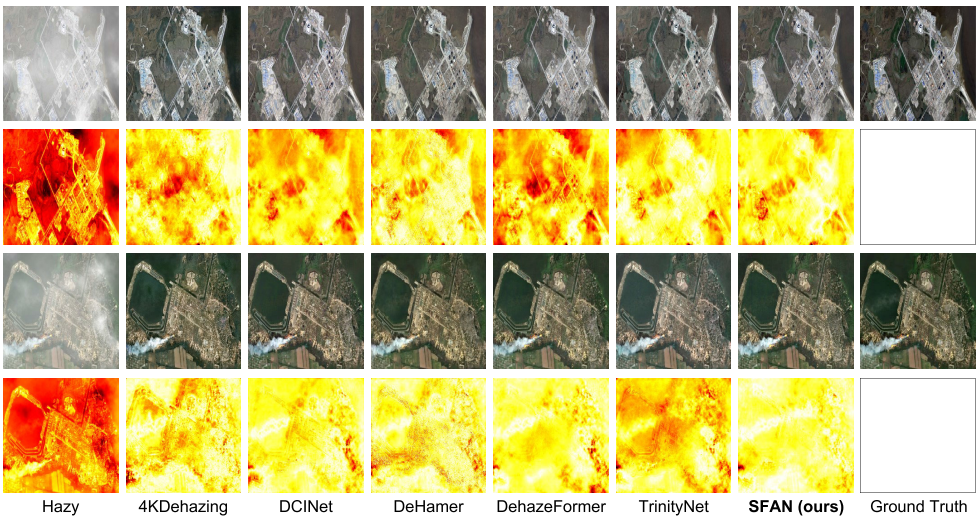

Fig. 8. Visual results and error maps on the RSID dataset [11]. The error map is generated by computing the difference between the dehazed image produced by each method and the corresponding ground-truth image.

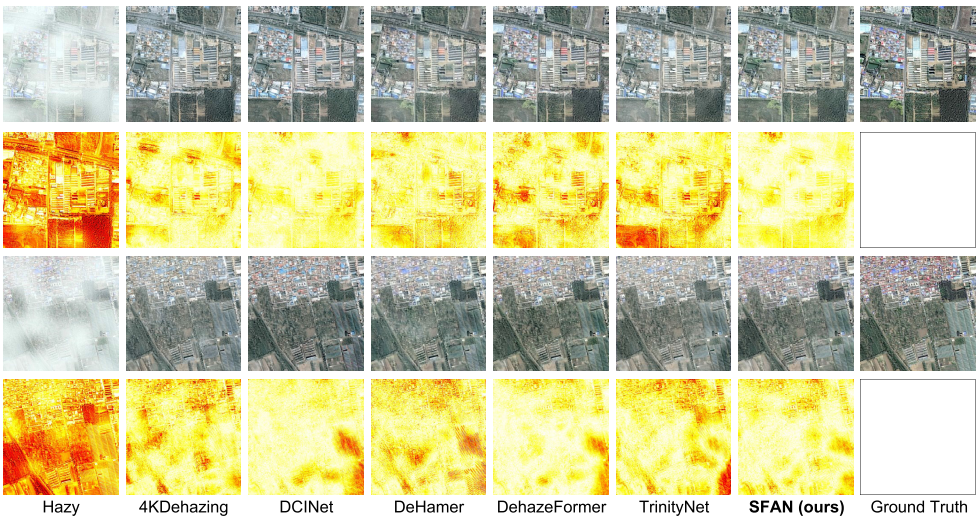

Fig. 9. Visual results and error maps on the SateHaze1K dataset [12]. The error map is generated by computing the difference between the dehazed image produced by each method and the corresponding ground-truth image.

attain superior performance in terms of PSNR and SSIM. Specifically, our method stands out across all benchmarks compared to recent RSI dehazing methods: TrinityNet [11] and PSMBNet [43]. For instance, on the DHID dataset,

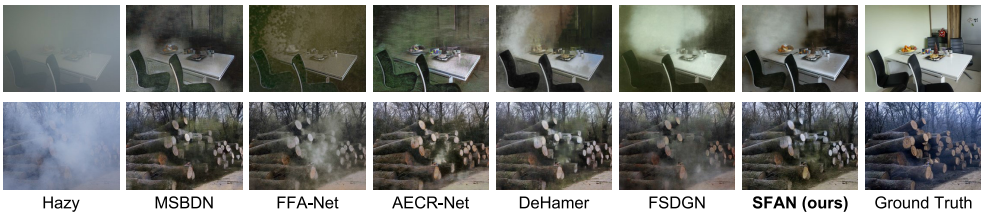

Fig. 10. Visual comparisons between our SFAN and the SOTA methods on the Dense-Haze and NH-HAZE [12].

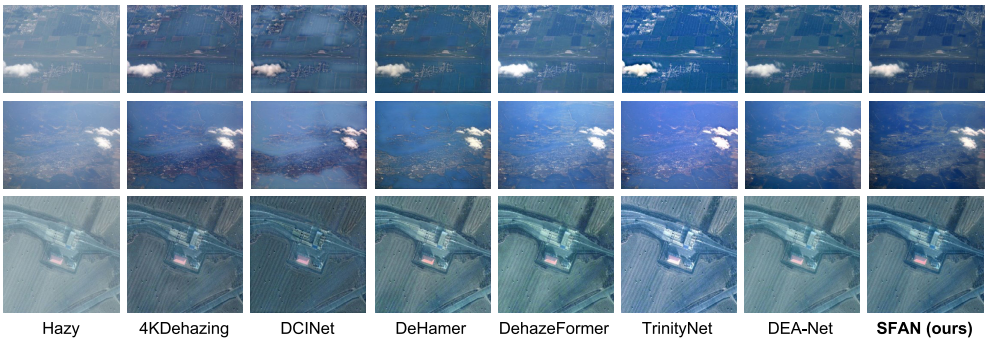

Fig. 11. Visual comparisons on the real-world remote sensing dataset (RRSD300 [62]).

TABLE II

QUANTITATIVE RESULTS OF REMOTE SENSING IMAGE DEHAZING ON SATEHAZE1K [12] DATASETS. THE BEST RESULTS ARE IN BOLD

Methods	S-thin [12]		S-moderate [12]		S-thick [12]	
	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑
SDCP [36]	0.825	14.186	0.786	16.068	0.842	16.385
MinVP [56]	0.937	20.967	0.913	21.046	0.863	16.588
IDeRs [57]	0.756	15.938	0.720	14.711	0.738	12.089
DCINet [10]	0.947	20.187	0.964	27.431	0.926	21.450
EMPFNet [18]	0.956	23.434	0.963	25.793	0.912	19.487
FCTFNet [19]	0.958	23.327	0.968	26.439	0.917	20.752
PSMBNet [43]	0.949	22.946	0.960	27.921	0.919	21.273
TrinityNet [11]	0.946	21.304	0.963	26.473	0.915	20.756
4KDehazing [58]	0.963	23.532	0.970	28.058	0.920	20.518
AECRNet [17]	**0.967**	**23.957**	0.958	26.091	0.924	21.457
DeHamer [20]	0.933	22.768	0.943	26.371	0.899	22.369
FSDGN [46]	0.946	21.730	0.970	25.771	0.937	22.351
MITNet [31]	0.950	22.054	0.963	23.637	0.918	20.190
DehazeFormer [22]	0.960	23.022	0.973	23.091	0.939	22.671
PhDnet-S [45]	0.960	23.266	0.971	26.854	0.937	22.257
DEA-Net [59]	0.961	23.512	0.973	26.971	0.938	22.983
SFAN (ours)	0.963	23.688	**0.977**	**28.191**	**0.942**	**23.006**

TABLE III

QUANTITATIVE RESULTS OF NATURAL IMAGE DEHAZING ON DENSE-HAZE [60] AND NH-HAZE [61]. THE BEST RESULTS ARE IN BOLD

Method	Venue	Dense-Haze [60]		NH-HAZE [61]	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DCP [6]	TPAMI'10	10.06	0.3854	10.57	0.5196
DehazeNet [13]	TIP'16	13.84	0.4252	16.62	0.5238
AOD-Net [63]	ICCV'17	13.14	0.4144	15.40	0.5693
GDN [64]	ICCV'19	13.31	0.3681	13.80	0.5370
FFA-Net [40]	AAAI'20	14.39	0.4524	19.87	0.6915
MSBDN [41]	CVPR'20	15.37	0.4858	19.23	0.7056
AECR-Net [17]	CVPR'21	14.88	0.5049	19.92	0.6717
DeHamer [20]	CVPR'22	16.62	0.5602	20.66	0.6844
FSDGN [46]	ECCV'22	16.91	0.5806	19.99	0.7086
DehazeFormer [22]	TIP'23	16.29	0.5100	20.47	0.7310
DEA-Net [59]	TIP'24	16.71	0.5578	20.51	0.7289
SFAN (ours)	-	**17.08**	**0.6128**	**20.69**	**0.7421**

TABLE IV

QUANTITATIVE RESULTS OF REAL-WORLD REMOTE SENSING IMAGE DEHAZING ON THE RRSD300 [62] DATASET

Methods	NIQE↓	FADE↓
Hazy	5.81	1.4351
4KDehazing [58]	5.60	0.8843
DCINet [10]	5.17	0.5487
DeHamer [20]	5.21	0.7268
DehazeFormer [22]	5.03	0.5591
TrinityNet [11]	5.12	0.5937
DEA-Net [59]	5.34	0.5570
SFAN (ours)	4.91	0.5239

SFAN outperforms them by 2.984 and 0.84 dB, respectively. Compared with the Transformer-based NSI dehazing methods, DeHamer [20] and DehazeFormer [22], SFAN achieves, on average, 5.52 and 0.23 dB higher PSNR results on the RICE dataset. Although DehazeFormer can achieve results close to ours on some datasets, it hardly achieves consistent improvements on all datasets. This is mainly attributed to the fact that our method adopts the MoE mechanism to dynamically modulate features, enabling the adaptiveness of the network for each sample. In addition, we compare the network parameters and FLOPs. As observed, although our method has more parameters compared to AECRNet, FSDGN, and MITNet, we achieve better performance with similar or lower FLOPs. For more comprehensive comparisons, we present the visual effects and the corresponding error maps in Figs. 6–8. As observed, our SFAN produces images that closely match the ground truth, as indicated by the corresponding error maps

(where lower brightness signifies closer alignment with the ground truth). Notably, our method excels at preserving structural integrity and accurately recovering details, as evidenced by the minimal errors produced.

Table II further compares the quantitative results on the SateHaze1K [12] dataset. As observed, our proposed SFAN can surpass the majority of methods on three haze levels. For instance, the AECR-Net [17] can perform best on SateHaze1K-thin but only obtains suboptimal results on other datasets. Although the thick haze results in severe degradation of image content, our method surpasses the Fourier-based method FSDGN [46] by 0.05 in SSIM. These results highlight the exceptional generalization capabilities of our method, allowing it to deliver outstanding performance across a variety of datasets. We also select two images to qualitatively analyze the performance of each method. Fig. 9 shows the visualized results of various methods, from which we can see that the difference between various methods is very small. However, by comparing the error maps, our SFAN still has slight superiority in removing haze.

C. Real-World Natural Image Dehazing

We further assess the performance of our proposed SFAN on real-world natural scene image dehazing datasets, such as Dense-Haze [60] and NH-HAZE [61]. The quantitative results are presented in Table III, showing that our SFAN outperforms all compared methods in terms of PSNR and SSIM. In detail, compared to the Transformer-based method Dehazeformer [22], our SFAN achieves noteworthy improvements of 0.79- and 0.22-dB PSNRs on these two datasets, respectively. This is primarily because our method not only fully leverages contextual information but also performs effective decoupled learning in the frequency domain. Furthermore, our method surpasses all metrics compared to frequency domain-based methods FSDGN [46], underscoring the efficacy of separate learning for low- and high-frequency features. The visual results of our SFAN and other compared methods are shown in Fig. 10, which also shows excellent performance. Since these two datasets contain thicker haze and do not have enough samples for training, our method cannot completely remove the haze degradation. However, other competitors perform worse than ours on these samples, with severe color distortion in the dehazed results.

D. Real-World Remote Sensing Image Dehazing

To evaluate our proposed SFAN on real-world remote sensing haze removal, we conduct comprehensive comparisons on the RRSD300 [62] dataset. We first present the naturalness image quality evaluator (NIQE [65]) and fog aware density evaluator (FADE [66]) scores that are associated with no-reference perceptual image quality in Table IV, from which we can see that our method gets the lower values, meaning a high-quality output with better perceptual results. The visual results can further be observed in Fig. 11, and our method can successfully eliminate haze and better restore image details, while 4KDehazing [58] and DCINet [10] fail to remove haze degradation. More importantly, our method does not cause a color shift or over-enhancement of dehazed images, which is better than TrinityNet [11].

TABLE V
ABLATION STUDY ON THE BLOCKS. FLOPS ARE COMPUTED ON THE IMAGE PATCH SIZE OF $256 \times 256 \times 3$

Model	MEB	DFLB	PSNR	Params.	FLOPs
Baseline	✗	✗	27.874	3.46M	15.92G
SFAN	✓	✗	28.527	3.40M	15.03G
	✗	✓	28.391	4.02M	17.25G
	✓	✓	29.173	3.95M	16.41G

TABLE VI
ABLATION STUDY ON THE **CROSS-SPATIAL** AND **CROSS-CHANNEL** MODULATION MECHANISMS. FLOPS ARE COMPUTED ON THE IMAGE PATCH SIZE OF $256 \times 256 \times 3$

Model	PSNR	Params.	FLOPs
(a) Cross-channel modulation	28.643	4.13M	16.48G
(b) Cross-spatial modulation	28.812	3.89M	16.14G
(c) Both	29.173	3.95M	16.41G

E. Ablation Studies

We conduct comprehensive ablation studies to demonstrate the effectiveness of our proposed components and the rationality of our design. Unless otherwise specified, all experiments are verified on the DHID [10] dataset according to the experimental settings.

1) Architecture Contribution: As detailed in Table V, we evaluate the effectiveness of our proposed key architectural components by benchmarking them against a baseline model. The baseline is implemented by replacing the MoME with three fixed modulation experts and employing the DFLB without decoupled learning for low- and high-frequency features, maintaining similar parameters and computational complexity. Incorporating our proposed modules into the baseline model results in significant and consistent improvements. Experimentally, the MEB achieves a performance gain of 0.653 dB in PSNR, while the DFLB provides a 0.517-dB improvement with only a marginal increase in parameters and FLOPs. When both modules are integrated, the model surpasses the baseline by 1.299 dB in PSNR, demonstrating the combined power of the proposed components. Notably, the final model achieves these gains without substantially increasing parameters and computational burden, significantly outperforming other methods. These results validate the efficacy of our design.

2) Design Choices of MoME: First, we investigate the importance of cross-spatial and cross-channel modulation mechanisms. Here, we fix the search space containing six modulation experts and sparsely select the top three experts to build the MoME. The cross-channel modulation model uses six different channel compression rates R, namely, $R = 1$, 2, 4, 8, 16, and 32. The cross-spatial modulation model uses three experts on the full-resolution and downsampled features, respectively. Due to the dynamicity of the network, the network parameters and model complexity of the three group experiments are slightly different. Based on the observed

TABLE VII

ABLATION STUDY ON THE NUMBER OF EXPERTS. E6K3 MEANS THAT THE MODULE CONTAINS SIX EXPERTS AND SPARSELY SELECTS THE TOP THREE EXPERTS FOR AGGREGATION. S-THICK IS THE ABBREVIATION OF THE SATE1K-THICK DATASET. FLOPS ARE COMPUTED ON THE IMAGE PATCH SIZE OF 256 × 256 × 3

Model	DHID	RSID	RICE1	S-thick	Params.	FLOPs
E4K3	28.891	25.886	36.501	22.325	4.03M	17.23G
E5K3	29.011	25.943	36.599	22.706	4.06M	16.81G
E6K3	**29.173**	**26.135**	36.653	**23.006**	3.95M	16.41G
E7K3	28.988	26.031	36.576	22.568	3.90M	16.05G
E8K3	28.766	25.980	36.526	22.512	3.95M	16.96G
E6K1	28.801	25.787	35.391	21.912	3.77M	14.95G
E6K2	29.085	26.021	36.453	22.612	3.83M	15.34G
E6K4	29.121	26.034	**36.888**	22.294	4.09M	16.93G

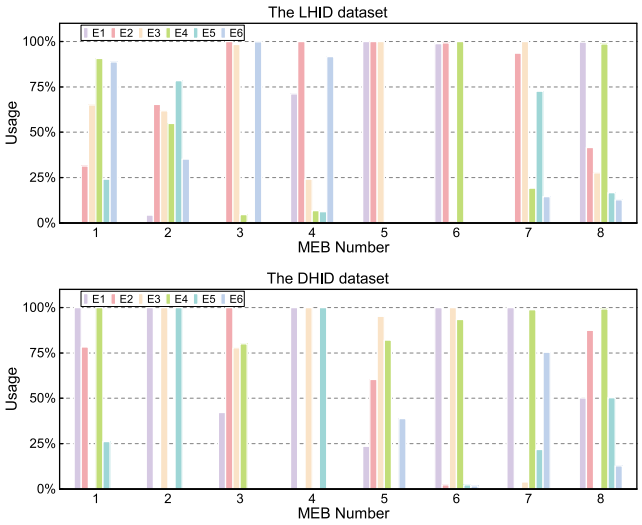

Fig. 12. Analysis for MEB. We plot the decisions made by the gating networks with the increasing network depth (from the first to the eight MoMB) on the LHID and DHID datasets [10].

TABLE VIII

ABLATION STUDY ON THE DFLB. THE ME AND SL REFER TO ME AND SL. SL-INV MEANS THAT WE PERFORM CONVOLUTION OPERATIONS ON THE PHASE SPECTRUM OF LF FEATURES AND THE AMPLITUDE SPECTRUM OF HIGH-FREQUENCY FEATURES. FLOPS ARE COMPUTED ON THE IMAGE PATCH SIZE OF 256 × 256 × 3

ME	SL	SL-Inv	MoFE	PSNR (dB)	Params.	FLOPs
✗	✗	✗	✗	28.227	3.31M	14.52G
✓	✗	✗	✗	28.831	3.36M	14.56G
✓	✓	✗	✗	29.021	3.76M	16.32G
✓	✗	✓	✗	28.544	3.78M	16.39G
✓	✓	✗	✓	29.173	3.95M	16.41G

results from Table VI, it is evidence that our hybrid modulation design yields the best performance. We speculate that as the depth of the network increases, the scale of features is constantly changing, and the requirements for feature reconstruction at each layer are also different. For instance, the lower layers often need more local information, while higher layers desire more global information. Therefore, our spatial and channel solution expands the search space, improves the model's adaptability to each layer, and further promotes performance improvements.

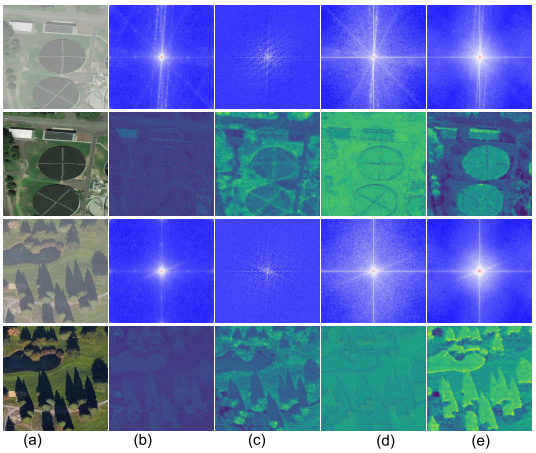

Fig. 13. Feature visualization in spatial and frequency domains. We show the features from the last DFLB in our SFAN. (a) Sampled images. (b) Input features. (c) LF features. (d) High-frequency (HF) features. (e) Fused features.

TABLE IX

ABLATION STUDY ON THE LOSS FUNCTIONS

Loss Function	DHID	RSID	RICE1	S-thick
$\mathcal{L}_{spa}$	29.007	25.782	35.413	22.451
$\mathcal{L}_{fre}$	27.326	23.312	34.671	21.349
$\mathcal{L}_{spa} + \alpha\mathcal{L}_{fre}$ (α=0.05)	**29.173**	**26.135**	**36.653**	**23.006**
$\mathcal{L}_{spa} + \alpha\mathcal{L}_{fre}$ (α=0.01)	29.195	25.941	36.607	22.502
$\mathcal{L}_{spa} + \alpha\mathcal{L}_{fre}$ (α=0.10)	29.288	26.034	36.591	22.563

Second, we conduct five sets of experiments, E4K3, E5K3, E6K3, E7K3, and E8K3, to investigate the impact of the number of experts on dehazing performance. For instance, E6K3 indicates that the MoME contains six modulation experts and sparsely selects the top three experts for feature fusion. As shown in Table VII, E6K3 outperforms other configurations with nearly the same model complexity. Next, we examine the influence of sparse expert selection on performance, as seen in configurations E6K1, E6K2, and E6K4. The results reveal that E6K4 achieves competitive results, even surpassing our method on the RICE1 dataset. However, E6K3 achieves superior performance on most datasets with lower model complexity. Consequently, we adopt E6K3 as our configuration for each MoME in this work. For MoFE, we perform similar ablation studies and select E4K2 as our final configuration.

Fig. 12 visualizes the decision-making process of the gate in each MEB on DHID and LHID datasets. We find that the decisions made by the gates are different on different datasets. Since we sparsely select three experts in each MEB, we observe that most layers, except the first and last layers, tend to select fixed three experts. However, the choice of gates is more diverse in the first and last layers. Furthermore, we find that the second expert (E2) is used in almost every MEB on the LHID dataset, and the first expert (E1) also performs the same on the DHID dataset. These show the dynamicity of the proposed network, enabling our method network the flexibility to adapt to different samples.

3) Design Choices of DFLB: We mainly explore the contribution of the ME, the SL, and the MoFE. Table VIII summarizes the performance improvements associated with

each component. Implementing the ME alone yields approximately a 0.6-dB performance gain, indicating that feature decoupling learning is more effective than unified learning for frequency features. Moreover, as presented in the fourth row, the specific SL schema enhances performance to over 29 dB. Conversely, when the inverse operation (SL-Inv) is used, e.g., employing convolution operations on the phase spectrum of LF features and the amplitude spectrum of high-frequency features, the performance drops dramatically. These observations underscore the soundness of our design. Finally, the results in the last row demonstrate that the proposed DFLB enhances performance by 0.946 dB compared to the baseline model, with only a minor increase in computational overhead.

Fig. 13 presents the spatial features and frequency spectral features before and after utilizing the DFLB. As we all know, the LF features mainly include global shapes and structures of an object or a scene, but the high-frequency features are more related to edges and texture. Notably, this is very consistent with our results of (c) and (d), which vividly showcases the effectiveness of our design. After integrating the HF and LF features based on the proposed MoFE, the visualized features clearly outline the global structure and local textures.

4) Loss Functions: Most of the previous dehazing methods only adopt spatial loss to optimize the network, yet our SFAN adopts a dual-domain interactive structure. Therefore, we conduct ablated experiments to verify the effectiveness of dual losses. As shown in Table IX, we observe the sole use of frequency loss can lead to performance degradation. However, when both losses are used without regard to the weighting factor, the performance is improved obviously on all four datasets. Next, we further compare the performance changes when the weight factor is set to different values. From the third to fifth rows, when set to $\alpha = 0.05$, we find that the results on most datasets outperform the other two sets of experiments. Therefore, we finally employ a hybrid loss function $\mathcal{L}_{\text{spa}} + \mathcal{L}_{\text{fre}}$ ($\alpha = 0.05$) to train our all models.

V. Conclusion

This article proposes a novel SFAN for remote sensing image dehazing. The core design is the proposed MoME and the DFLB. Specifically, the MoME employs the MoE mechanism to adaptively extract rich cross-scale and cross-channel contextual features, which are then used for feature modulation. The DFLB processes signals from different frequency bands in the input hazy images, utilizing a dual-branch structure to implement decoupled learning of HF and LF features. This enables effective feature modulation and interaction between the hierarchical decoder features and frequency features. Experiments on various benchmark remote sensing hazy datasets demonstrate that our method surpasses recent state-of-the-art approaches. Although excellent performance improvements have been achieved, most current remote sensing dehazing methods train a separate model for each dataset. The next focus of our research will be on how to train a unified remote sensing dehazing model to adapt to different dehazed images. Furthermore, the other line will aim at enhancing real-world dehazing performance and addressing the complexities of remote sensing haze removal.

Acknowledgment

The computation is completed on the HPC Platform of Hefei University of Technology. The author Hao Shen deeply thanks her overseas supervisors, Xudong Jiang from the Department of Electrical and Electronic Engineering, Nanyang Technological University, for his valuable guidance and support on his research work.

References

[1] Y. Liu et al., "Improved techniques for learning to dehaze and beyond: A collective study," 2018, *arXiv:1807.00202*.
[2] W. Ren et al., "Deep video dehazing with semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1895–1908, Apr. 2019.
[3] B. Wang, Z. Zhang, S. Zhao, H. Zhang, R. Hong, and M. Wang, "CropCap: Embedding visual cross-partition dependency for image captioning," in *Proc. 31st ACM Int. Conf. Multimedia*, vol. 1409, Oct. 2023, pp. 1750–1758.
[4] K. Li, D. Guo, and M. Wang, "Proposal-free video grounding with contextual pyramid network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1902–1910.
[5] K. Li, D. Guo, and M. Wang, "ViGT: Proposal-free video grounding with a learnable token in the transformer," *Sci. China Inf. Sci.*, vol. 66, no. 10, Oct. 2023, Art. no. 202102.
[6] 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, Dec. 2010.
[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, Mar. 2020.
[8] A. Makarau, R. Richter, R. Müller, and P. Reinartz, "Haze detection and removal in remotely sensed multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5895–5905, Sep. 2014.
[9] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, "A remote sensing image dataset for cloud removal," 2019, *arXiv:1901.00600*.
[10] L. Zhang and S. Wang, "Dense haze removal based on dynamic collaborative inference learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5631016.
[11] 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, 2023, Art. no. 4702914.
[12] B. Huang, Z. Li, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1806–1813.
[13] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.
[14] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3194–3203.
[15] W. Ren, J. Pan, H. Zhang, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks with holistic edges," *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 240–259, Jan. 2020.
[16] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 6, pp. 713–724, Jun. 2003.
[17] H. Wu et al., "Contrastive learning for compact single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10551–10560.
[18] Y. Wen, T. Gao, J. Zhang, Z. Li, and T. Chen, "Encoder-free multiaxis physics-aware fusion network for remote sensing image dehazing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4705915.
[19] 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, Oct. 2021.
[20] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5812–5820.
[21] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, and Z. Jin, "MB-TaylorFormer: Multi-branch efficient transformer expanded by Taylor formula for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12802–12813.

[22] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Trans. Image Process.*, vol. 32, pp. 1927–1941, 2023.
[23] P. Dong and B. Wang, "TransRA: Transformer and residual attention fusion for single remote sensing image dehazing," *Multidimensional Syst. Signal Process.*, vol. 33, no. 4, pp. 1119–1138, Dec. 2022.
[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[25] X. Ma et al., "Efficient modulation for vision networks," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–19.
[26] J. Yang, C. Li, X. Dai, and J. Gao, "Focal modulation networks," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 4203–4217.
[27] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang, "TTST: A top-k token selective transformer for remote sensing image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 738–752, 2024.
[28] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: A literature survey," *Artif. Intell. Rev.*, vol. 42, pp. 275–293, Aug. 2014.
[29] J. Hou, Q. Cao, R. Ran, C. Liu, J. Li, and L.-J. Deng, "Bidomain modeling paradigm for pansharpening," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 347–357.
[30] Y. Xiao, Q. Yuan, K. Jiang, Y. Chen, Q. Zhang, and C.-W. Lin, "Frequency-assisted mamba for remote sensing image super-resolution," 2024, *arXiv:2405.04964*.
[31] H. Shen, Z.-Q. Zhao, Y. Zhang, and Z. Zhang, "Mutual information-driven triple interaction network for efficient image dehazing," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7–16.
[32] K. Jiang, J. Jiang, X. Liu, X. Xu, and X. Ma, "FMRNet: Image deraining via frequency mutual revision," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 11, pp. 12892–12900.
[33] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Image restoration via frequency selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1093–1108, Feb. 2024.
[34] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3522–3533, Nov. 2015.
[35] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1674–1682.
[36] J. Li, Q. Hu, and M. Ai, "Haze and thin cloud removal via sphere model improved dark channel prior," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 472–476, Mar. 2019.
[37] H. Shen, C. Zhang, H. Li, Q. Yuan, and L. Zhang, "A spatial–spectral adaptive haze removal method for visible remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6168–6180, Sep. 2020.
[38] Z. Gu, Z. Zhan, Q. Yuan, and L. Yan, "Single remote sensing image dehazing using *a prior*-based dense attentive network," *Remote Sens.*, vol. 11, no. 24, p. 3008, Dec. 2019.
[39] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, and M.-H. Liu, "Gated fusion network for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3253–3261.
[40] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Xie, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 11908–11915.
[41] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2157–2167.
[42] Y. Zhang, S. Zhou, and H. Li, "Depth information assisted collaborative mutual promotion network for single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 2846–2855.
[43] H. Sun et al., "Partial Siamese with multiscale bi-codec networks for remote sensing image haze removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4106516.
[44] Y. Du, J. Li, Q. Sheng, Y. Zhu, B. Wang, and X. Ling, "Dehazing network: Asymmetric UNet based on physical model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.
[45] Z. Lihe, J. He, Q. Yuan, X. Jin, Y. Xiao, and L. Zhang, "PhDNet: A novel physic-aware dehazing network for remote sensing images," *Inf. Fusion*, vol. 106, Jun. 2024, Art. no. 102277.
[46] 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.*, Oct. 2022, pp. 181–198.
[47] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 120, pp. 1–39, 2022.
[48] D. Lepikhin et al., "GShard: Scaling giant models with conditional computation and automatic sharding," 2020, *arXiv:2006.16668*.
[49] X. He, K. Yan, R. Li, C. Xie, J. Zhang, and M. Zhou, "Frequency-adaptive pan-sharpening with mixture of experts," 2024, *arXiv:2401.02151*.
[50] B. Cao, Y. Sun, P. Zhu, and Q. Hu, "Multi-modal gated mixture of local-to-global experts for dynamic image fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 23555–23564.
[51] H. Yang, L. Pan, Y. Yang, and W. Liang, "Language-driven all-in-one adverse weather removal," 2023, *arXiv:2312.01381*.
[52] J. Hou et al., "Linearly-evolved transformer for pan-sharpening," 2024, *arXiv:2404.12804*.
[53] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 12312–12321.
[54] M. Zhou et al., "Adaptively learning low-high frequency information integration for pan-sharpening," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3375–3384.
[55] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.
[56] 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.
[57] 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, Oct. 2017.
[58] Z. Zheng et al., "Ultra-high-definition image dehazing via multi-guided bilateral learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16180–16189.
[59] 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, 2024.
[60] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-Haze: A benchmark for image dehazing with dense-haze and haze-free images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1014–1018.
[61] C. O. Ancuti, C. Ancuti, and R. Timofte, "NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 444–445.
[62] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.
[63] 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.*, Jun. 2017, pp. 4770–4778.
[64] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7314–7323.
[65] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Apr. 2012.
[66] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3888–3901, Nov. 2015.

Hao Shen is currently pursuing the Ph.D. degree with the School of Computer and Information, Hefei University of Technology, Hefei, China.

He has published articles in ACM MM, AAAI, CVPR, ECAI, TGRS, and PR. His research interests include image processing, computer vision, and deep learning.

Henghui Ding (Member, IEEE) received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 2016. He received the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2020.

He was a Research Scientist at ByteDance, Beijing, China, and a Post-Doctoral Researcher at ETH and NTU. He is currently a tenure-track Professor at Fudan University, Shanghai, China. He serves as an Associate Editors for IET Computer Vision and Visual Intelligence and serves/has served as the Area Chair of CVPR'24, NeurIPS'24, ICLR'25, ACM MM'24, and BMVC'24 and an SPC Member of AAAI'(22–25) and IJCAI'(23–24). His research interests include computer vision and machine learning.

Yulun Zhang (Member, IEEE) received the B.E. degree from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2013, the M.E. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2017, and the Ph.D. degree from the Department of ECE, Northeastern University, Boston, MA, USA, in 2021.

He is an Associate Professor at Shanghai Jiao Tong University, Shanghai, China. He was a Post-Doctoral Researcher at Computer Vision Laboratory, ETH Zürich, Zürich, Switzerland. His research interests include image/video restoration and synthesis, biomedical image analysis, model compression, multimodal computing, large language models, and computational imaging.

Dr. Zhang is/was the Area Chair of CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, IJCAI, and ACM MM and a Senior Program Committee (SPC) Member of IJCAI and AAAI.

Xiaofeng Cong (Student Member, IEEE) received the B.S. and M.S. degrees from the School of Electrical Engineering and Automation, Anhui University, Hefei, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree with the School of Cyber Science and Engineering, Southeast University, Nanjing, China.

He has published papers in ACM MM, ACM CSUR, TMM, and IJCAI. His research interests include image dehazing and generative adversarial networks.

Zhong-Qiu Zhao (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2007.

From 2008 to 2009, he held a postdoctoral position in image processing at the CNRS UMR6168 Laboratory Sciences de Information et des Systmes, La Garde, France. From 2013 to 2014, he was a Research Fellow in image processing with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. He is currently a Professor with Hefei University of Technology, Hefei. His research interests include pattern recognition, image processing, and computer vision.

Xudong Jiang (Fellow, IEEE) received the B.E. and M.E. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1983 and 1986, respectively, and the Ph.D. degree from Helmut Schmidt University, Hamburg, Germany, in 1997.

From 1998 to 2004, he was with the Institute for Infocomm Research, A*STAR, Singapore, as a Lead Scientist, and the Head of the Biometrics Laboratory. He joined Nanyang Technological University (NTU), Singapore, as a Faculty Member, in 2004, where he was the Director of the Center for Information Security from 2005 to 2011. He is currently a Professor with the School of EEE, NTU, and the Director of the Centre for Information Sciences and Systems, Singapore. He has authored more than 200 articles with over 60 articles in IEEE journals and 30 papers in top conferences such as CVPR/ICCV/ECCV/AAAI/ICLR. His research interests include computer vision, machine learning, pattern recognition, image processing, and biometrics.

Dr. Jiang has served as an Associate Editor for IEEE SIGNAL PROCESSING LETTERS (IEEE SPL) and IEEE TRANSACTIONS ON IMAGE PROCESSING (IEEE T-IP). Currently, he serves as a Senior Area Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and the Editor-in-Chief of *IET Biometrics*.