Spatial Frequency Modulation Network for Efficient Image Dehazing

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

***Abstract*—Currently, two main research lines in efficient context modeling for image dehazing are tailoring effective feature modulation mechanisms and utilizing the Fourier transform more precisely. The former is usually based on self-scale features that ignore complementary cross-scale/level features, and the latter tends to overlook regions with pronounced haze degradation and intricate structures. This paper introduces a novel spatial and frequency modulation perspective to synergistically investigate contextual feature modeling for efficient image dehazing. Specifically, we delicately develop a Spatial Frequency Modulator (SFM) equipped with a Cross-Scale Modulator (CSM) and Frequency Modulator (FM) to implement intra-block feature modulation. The CSM progressively aggregates hierarchical features across different scales, employing them for spatial self-modulation, and the FM subsequently adopts a dual-branch design to focus more on the crucial areas with severe haze and complex structures for reconstruction. Further, we propose a Cross-Level Modulator (CLM) to facilitate inter-block feature mutual modulation, enhancing seamless interaction between features at different depths and layers. Integrating the above-developed modules into the U-Net architecture, we construct a two-stage spatial frequency modulation network (SFMN). Extensive quantitative and qualitative evaluations showcase the superior performance and efficiency of the proposed SFMN over recent state-of-the-art image dehazing methods. The source code can be found in https://github.com/it-hao/SFMN.**

***Index Terms*—Image dehazing, spatial frequency modulation, cross-scale, cross-level.**

Received 11 July 2024; revised 15 April 2025 and 3 June 2025; accepted 6 June 2025. Date of publication 19 June 2025; date of current version 25 June 2025. This work was supported in part by the Scientific Research Foundation for High-Level Talents of Anhui University of Science and Technology under Grant 2025yjrc0015, in part by the National Natural Science Foundation of China (NSFC) under Grant 62472104, and in part by Anhui Special Support Plan for High-Level Talents and Guangxi Key Research and Development Program under Grant AB22035022. The associate editor coordinating the review of this article and approving it for publication was Dr. Lizhi Wang. *(Corresponding author: Zhong-Qiu Zhao.)*

Hao Shen is with the School of Public Security and Emergency Management, Anhui University of Science and Technology, Hefei 231131, China, also with the State Key Laboratory of Digital Intelligent Technology for Unmanned Coal Mining, Anhui University of Science and Technology, Huainan 232001, China, and also with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: haoshenhs@gmail.com).

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 AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yulun100@gmail.com).

Zhong-Qiu Zhao is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: z.zhao@hfut.edu.cn).

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

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2025.3579148, provided by the authors.

Digital Object Identifier 10.1109/TIP.2025.3579148

I. Introduction

THE images captured in hazy or adverse scenes significantly influence the performance of high-level vision tasks, such as object detection [1] and scene understanding [2], [3], [4]. Therefore, image dehazing, which aims to reconstruct haze-free images from their corresponding hazy images, has been a hot topic in academic and industry communities.

Over the past decade, inspired by the success of deep learning, convolutional neural network (CNN)-based methods [6], [7], [8], [9], [10], [11], [12] have achieved superior performance. Among them, early approaches [6], [11], [12] focus on estimating the parameters of the atmosphere scattering model [13] by CNNs. Lately, more image dehazing approaches [9], [14], [15], [16] employ an end-to-end network to estimate clear, hazy-free images directly. Nevertheless, due to the local modeling properties of CNNs, these methods still have limited abilities to capture long-range dependencies critical for image dehazing. Recently, Transformer-based methods [17], [18], [19], [20] have displayed excellent global context modeling capabilities in low-level vision tasks, including image dehazing, mainly designed based on spatial or channel self-attention mechanisms. However, for self-attention in spatial dimensions, its computational complexity increases quadratically with the resolution of the feature map. Channel-based self-attention cannot model spatial long-range dependencies well, thus limiting performance improvements. We aim to construct a purely CNN-based dehazing network that embraces the merits of convolution and self-attention mechanisms, namely having fine-grained detail reconstruction and mining global feature dependencies.

FocalNet [5] first gathers contexts around each query and subsequently modulates the query using the generated context (Fig. 1 (a)), which decouples the aggregation from the individual queries, making the interactions between features more lightweight and efficient, which is essential for building efficient dehazing networks. The element-wise multiplication operation used in FocalNet can project features into an extremely high-dimensional implicit feature space with lower computational complexity [22]. By stacking multiple layers of this operation in the network, the implicit dimension can be increased exponentially to near infinity in a recursive manner. The matrix multiplication in self-attention shares similar attributes (non-linearity and high dimensionality) with element-wise multiplication. However, this method neglects to exploit cross-scale feature modulation (Fig. 1 (b)), whereas

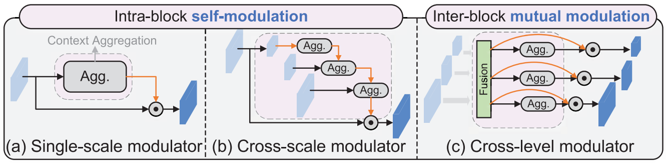

Intra-block self-modulation
Inter-block mutual modulation
Context Aggregation
Agg.
Agg.
Agg.
Agg.
Fusion
Agg.
Agg.
Agg.
(a) Single-scale modulator
(b) Cross-scale modulator
(c) Cross-level modulator

Fig. 1. Comparisons between previous feature modulation rule (single-scale modulator [5]) and our proposed cross-scale and cross-level modulators.

cross-scale design has receptive fields of different sizes, which is more suitable for adapting to the uneven distribution of haze, thus, cross-scale representation learning helps remove haze degradation at different scales. Meanwhile, high-resolution features encompass more geometric details but may lack awareness of contextual information, while low-resolution features exhibit the exact opposite characteristics. In addition, features at different layers encode variant information at different scales [23], [24]. Therefore, utilizing a mutual-modulated manner to perform cross-level feature interaction (Fig. 1 (c)) may benefit representation learning.

The other research line of context feature modeling is implemented by applying Fourier transform [16], [25], [26], [27]. Benefiting from the innate global properties of the Fourier domain, several works [16], [25], [28], [29] propose to utilize frequency statistical information to guide image restoration. Most of them perform frequency learning on the spatial dimension and have explored that the degradation property induced by haze primarily manifests in the amplitude spectrum [16], [28]. However, these methods pay less attention to informative signals, such as edges or regions that contain severe haze degradation.

Motivated by the above-mentioned analysis, we propose an effective and computationally efficient two-stage network named Spatial Frequency Modulation Network (SFMN) for image dehazing, which follows the trend of feature modulation but from spatial and frequency dual domains. Firstly, towards intra-block feature modulation, we formulate a *spatial frequency modulator* (SFM), which comprises a *cross-scale modulator* (CSM), expertized in modeling spatial contextual relationships, and a *frequency modulator* (FM), focus on removing global haze degradation and reconstructing imperative regions. Technically, the CSM adopts a multi-branch structure to progressively aggregate cross-scale hierarchical context, and then utilizes element-wise multiplication to facilitate feature interaction between query features and aggregated context. The FM firstly leverages the Fourier domain's global properties to channel-dependent plain features and channel-independent discriminative features separately and then leaves the latter to modulate the former to enhance crucial regions selectively. It is worth emphasizing that we deal with Fourier amplitude and phase information separately in the two-stage networks to facilitate learning of the frequency-specific information and reduce the network's optimization difficulty.

Furthermore, we design a *cross-level modulator* (CLM) to realize inter-block feature mutual modulation in the encoder and decoder. In detail, it firstly leverages inherent cross-level features from the encoder-decoder for both top-down and bottom-up information flow, then employs a gating mechanism to achieve context aggregation, and finally adopts the element-wise multiplication to complete feature mutual modulation.

Different from the existing literature [5], [8], [16], [28], the proposed method fully couples the cross-scale and cross-level features in the spatial domain with the crucial features in the frequency domain by element-wise multiplication operation, thus achieving performance superior to Transformer-based methods with lower computational complexity. As depicted in Fig. 2, our network significantly outperforms the state-of-the-art (SOTA) Transformer-based models such as MBTFormer-B [20] and DeHamer [17] while utilizing only half or even less of the FLOPs. Compared with efficient models FSDGN [16] and AECR-Net [9], our network can strike a better trade-off between performance and model complexity.

The main contributions of this study are listed as follows:

- In the spatial domain, we design a cross-scale modulator and cross-level modulator based on the element-wise multiplication operation, which realizes self-modulation of intra-block features and mutual modulation of inter-block features, respectively, and improves high-dimensional representational capacity with lower computational complexity.
- In the frequency domain, we design a frequency modulator and combine it with the cross-scale modulator to build the spatial frequency modulator, which serves as the basic module of the subsequent network.
- Based on the spatial frequency modulator, we propose SFMN, a two-stage network incorporating dual-domain modulation in both the spatial and frequency domains, to achieve a more efficient and accurate image dehazing.
- We perform comprehensive experiments on multiple benchmark datasets, mainly focusing on dehazing tasks, to showcase the superior performance of our SFMN compared to state-of-the-art methods.

II. Related Work

A. Single Image Dehazing

There are many representative dehazing algorithms, mainly including prior-based, data-driven-based, and neural augmentation-based methods.

Prior-based methods are the trailblazers in image dehazing, typically relying on atmospheric scattering models (ASM) [13] and handcrafted priors. Notable examples of these priors include the dark channel prior (DCP) [30], non-local prior (NLP) [31], and color attenuation prior (CAP) [32]. He et al. [30] are the first to propose DCP combined with the haze imaging model to estimate haze-free images directly. Li et al. [33] introduce an adaptive sky compensation term to address noise amplification in the sky region, thereby reducing morphological artifacts. Subsequently, Zhu et al. [32] utilize a linear model to determine the scene depth of hazy images based on the CAP, enabling haze removal through the obtained depth information. Berman et al. [31] propose a non-local dehazing method, leveraging the observation that pixels within a given cluster are distributed across the entire image and

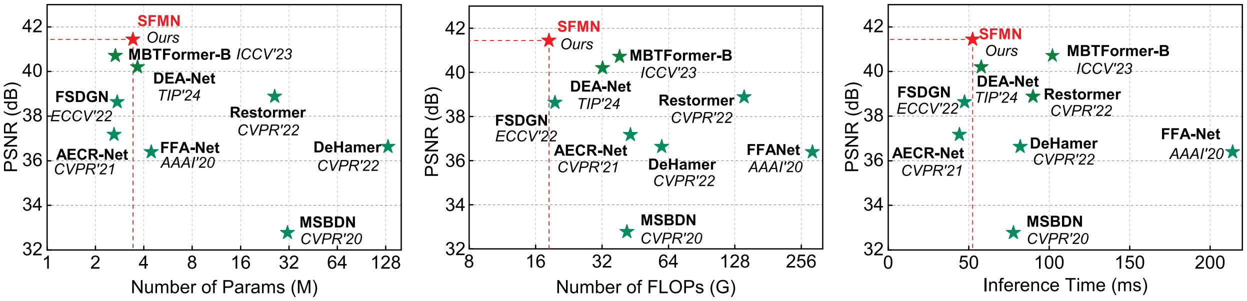


Fig. 2. Model complexity comparisons of SOTA methods, where the FLOPs are calculated with the image patch of 256 × 256, and the average inference time is calculated on SOTS-Indoor [21] by repeating experiments five times.

located at varying distances from the camera. Li et al. [34] further construct a dark direct attenuation prior (DDAP) to solve the ambiguity between object radiance and the haze and noise amplification in sky regions, achieving superior subjective quality evaluation. However, these prior-based methods may lead to transmission estimation errors due to imprecise prior information, producing visually unpleasant dehazing images with slight color distortion or other degradation phenomena. Recent research [9], [14], [35], [36] has focused on directly restoring clean images from the corresponding hazy images, eliminating the necessity for a physical model. Among them, GFN [8] introduces a groundbreaking fusion-based strategy to generate haze-free images. FFA-Net [14] introduces a feature attention block constructed with channel attention and pixel attention mechanisms capable of handling diverse forms of hazy images. MSBDN [35] leverages the boosting strategy and the back-projection technique to achieve dense feature fusion. AECR-Net [9] proposes contrast regularization based on contrastive learning to restore hazy images to haze-free images close to clean images, achieving commendable trade-offs between network parameters and performance. Zheng et al. propose a series of ultra-high-definition image dehazing methods [36], [37], [38] to address the key challenges of slow training speed and high memory consumption. C2PNet [39] further proposes a curricular contrastive regularization mechanism and a physics-aware dual-branch unit for image dehazing. DIACMPN [40] integrates depth estimation and dehazing by a dual-task interaction mechanism and achieves mutual enhancement of performance. However, all these methods rarely explore long-range dependencies due to the intrinsic locality properties of CNNs, thus resulting in suboptimal performance. Subsequently, the self-attention-based Transformers have been introduced into image dehazing to model long-range dependencies and significantly improve the performance. DeHamer [17] brings density-related priors into the Transformer architecture and combines CNNs to achieve local and global representations. DehazeFormer [19] redesigns a critical structure of the Swin Transformer [41] to better suit the task of image dehazing. MBTFormer [20] utilizes the Taylor expansion and multi-scale patch embedding to construct a network with linear computational complexity. DEA-Net [42] proposes to use detail-enhanced convolution to replace vanilla convolution and designs a content-guided attention mechanism to handle haze non-uniformity. However, these methods operate exclusively in the spatial domain, ignoring the exploration of context modeling in the frequency domain and the characteristics of haze degradation.

In addition, neural augmentation [43] -based methods combining priors and CNNs have also been gradually proposed. Zhao et al. [44] propose the RefineDNet, which adopts prior-based DCP to restore visibility and then employs GANs to improve realness. Li et al. [45] first estimate transmission maps and atmospheric light and then adopt dual-scale GANs to refine the results. Although these methods cannot achieve end-to-end dehazing, they can achieve superior performance on synthesized and real-world hazy images.

B. Frequency Learning in Low-Level Vision

In low-level tasks, high-frequency signals usually refer to image details and textures, whereas low-frequency signals represent flat regions. Both of them are important for reconstructing the restored images. There are two main types of frequency-based methods: wavelet transform and Fourier transform. SDWNet [46] proposes a wavelet reconstruction module to enrich high-frequency features for image deblurring. MWCNN [47] designs a novel multi-level wavelet network to recover detailed textures and sharp structures. WSAMF-Net [48] builds wavelet spatial attention to enhance the extracted features for better structures and edges.

Recently, the effectiveness of the Fourier transform in global modeling for low-level tasks has been demonstrated by various studies [16], [25], [49], [50]. DeepRFT [51] incorporates ReLU in the frequency domain to extract kernel-level information, seamlessly integrating it into the ResBlock for effective deblurring. FSDGN [16] pioneers the revelation that the degradation property of hazy images primarily manifests in the amplitude spectrum and designs a frequency and spatial dual guidance network, with similar designs found in [49]. MITNet [28] utilizes a mutual-information constraint to achieve spatial-frequency feature complementary learning. This study focuses more on learning from the important regions with severe haze degradation and complex structures.

III. Methods

A. Overall Architecture

As shown in Fig. 3, we adopt a two-stage design, and each is constructed based on the U-Net [52]. *This design aims to*

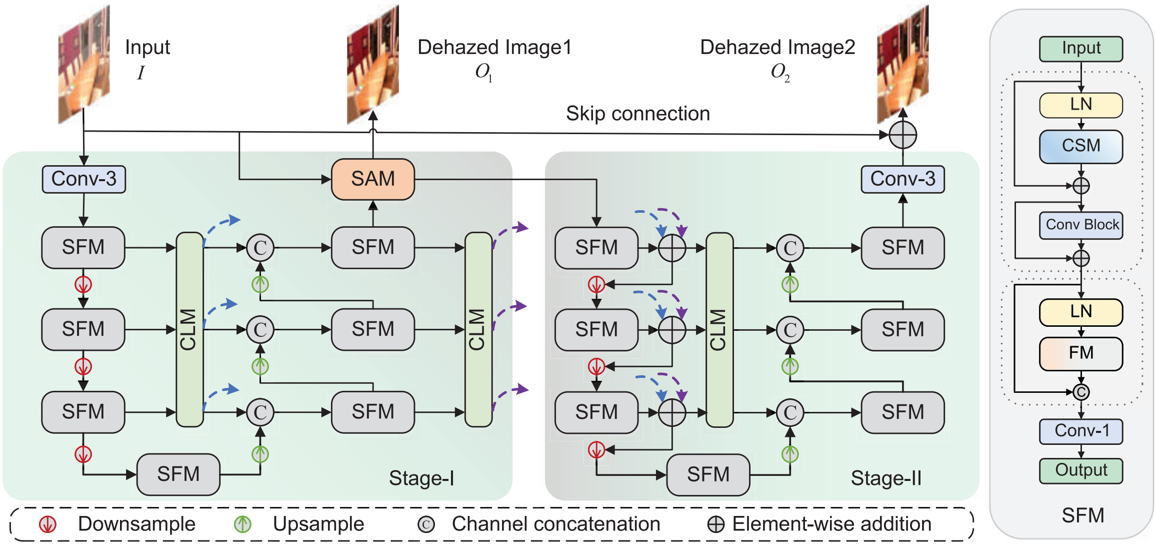


Fig. 3. The overall architecture of our proposed spatial frequency modulation network (SFMN). We adopt a two-stage U-Net design to construct the whole network, where the proposed spatial frequency modulator (SFM) is used as the basic building block in each layer. The encoder in the second stage incorporates the corresponding level's output of the CLM from the encoder (blue dashed line) and decoder (purple dashed line) of the first stage network.

disentangle the Fourier prior information, e.g., amplitude and phase spectrums, according to their unique characteristics and then process them independently in each stage. In the first stage, given a hazy image $I \in \mathbb{R}^{H\times W\times 3}$, the SFMN firstly adopts a 3×3 convolution to process it and obtain the shallow features I_0. Next, the shallow features are sent to a four-level symmetric encoder-decoder. Each encoder or decoder contains a spatial frequency modulator (SFM). Give I_0 as input, and encoders decrease the spatial resolution by half while increasing the number of feature channels as the stage progresses. On the other hand, decoders, using low-resolution features as input, decrease half of the feature channels while doubling the size of feature maps. A cross-level modulator (CLM) used in the encoder is designed to implement multi-scale feature interaction and intra-block global feature modeling. Then, the upsampled features are concatenated with the output of the corresponding level from the CLM to improve the information flow. In the end, following [53], we introduce the supervised attention module (SAM) to generate output images, represented as O_1, allowing useful ones to propagate to the next stage and making stable optimization.

Like the first stage, the second stage adopts almost the same structure to extract features. However, several crucial points warrant emphasis. **(i)** Instead of directly using I as input, we opt for combining the phase spectrum of the input image and the amplitude spectrum of the first stage's output image to function as the input features. **(ii)** The FM in the second stage performs the Fourier phase spectrum learning rather than the amplitude spectrum learning. **(iii)** The encoder incorporates the corresponding level's output of the CLM (purple and blue dashed lines in the Fig. 3) from the first stage network, which enhances feature propagation, preserves the fine structural details in the original images, and ensures stable network training. **(iv)** A convolutional layer is applied to generate the residual image, which is added to the original hazy image to obtain the final output O_2.

B. Spatial Frequency Modulator (SFM)

Unlike traditional Transformer [54] architectures, our proposed SFM operates on dual-domain features. As shown in Fig. 4, it mainly consists of two LayerNorm (LN), a cross-scale modulator (CSM), a convolutional block, and a frequency modulator (FM). By synergistically organizing them, the network can effectively realize contextual self-modulation of intra-block features.

1) Cross-Scale Modulator: As displayed in Fig. 4 (a), to achieve cross-scale spatial feature aggregation, we adopt a multi-branch architecture to achieve spatial context extraction of specific-scale features and exploit them in a progressive strategy. Supposing the input features as $F_{cs} \in \mathbb{R}^{H\times W\times C}$, we first utilize average pooling (AP) operations with different downsampling ratios to convert F_{cs} into distinct scale-spaces. For each branch, the resulting features obtained by passing through HFE are incorporated into the next branch via feature upsampling followed by an element-wise addition operation. Therefore, the design is capable of removing haze information in a coarse-to-fine manner. Formally, the whole operation process in the i-th branch can be expressed as:

$$M_{cs,i} = \mathrm{HFE}(\mathrm{AP}_{2^{b-i}}(F_{cs}) + \mathrm{UP}_2(M_{cs,i-1})), \quad (1)$$

where b denotes the number of branches, $\mathrm{AP}_{2^{b-i}}$ ($i = 1, 2, \cdots, b$) denotes the average pooling with the downsampling rate 2^{b-i}, UP_2 denotes the interpolation upsampling with the upsampling rate as 2, and $\mathrm{HFE}(\cdot)$ is the hierarchical feature extractor function. $M_{cs,b}$ is the output of the last branch and $M_{cs,0} = 0$.

For hierarchical feature extractor in the i-th branch, we stack $L = 3$ depth-wise convolutions (DConvs) with different kernel sizes to progressively extract contextualization features. Specifically, we initialize the kernel size with 3×3 and gradually increase it by two per layer. As for the l-th layer,

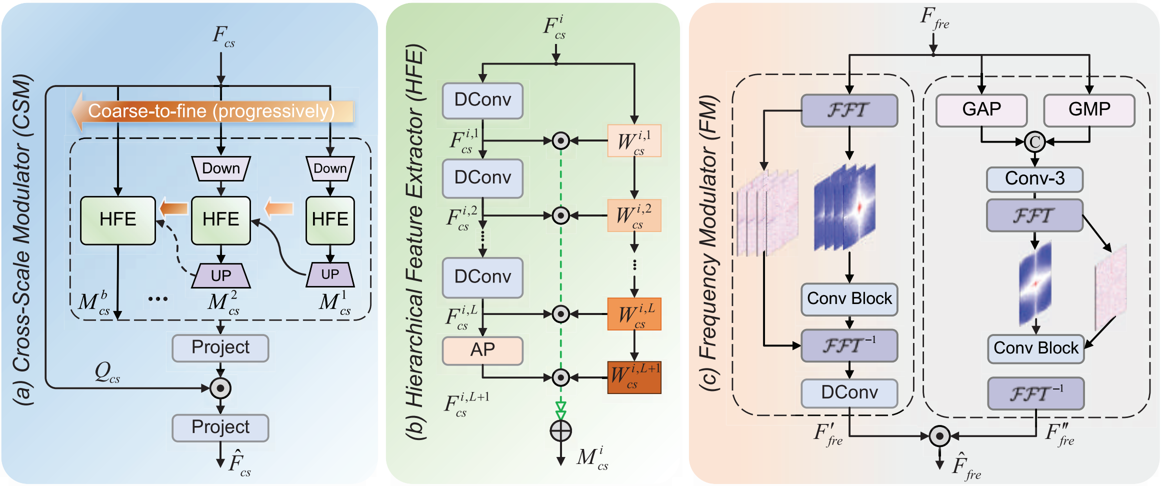


Fig. 4. Overview of the core components. (a) The architecture of the proposed cross-scale modulator (CSM). (b) The architecture of the hierarchical feature extractor (HFE). (c) The architecture of the proposed frequency modulator (FM).

the whole operation process can be denoted as

$$F^{i,l}_{cs} = \mathrm{GELU}(\mathrm{DConv}_{k\times k}(F^{i,l-1}_{cs})), \tag{2}$$

where $F^{i,l}_{cs}$ denotes the output features in l-th layer of the i-th branch. This series of operations enables the network to obtain multi-granularity information. Then, we perform a global average pooling on $F^{i,L}_{cs}$ to obtain global spatial information. Thus, there are $L+1$ feature maps in total. Subsequently, we employ a *gating mechanism* to interact between query and multi-granularity features to dynamically amalgamate the obtained coarse-to-fine features. In detail, the element-wise multiplication is initially applied to the respective gating weights and features. Next, we utilize element-wise addition to aggregate all modulated features. The calculation procedure can be formulated as:

$$M^i_{cs} = \sum_{\ell=1}^{L+1} W^{i,l}_{cs} \odot F^{i,l}_{cs}, \tag{3}$$

where $W^{i,l}_{cs} \in \mathbb{R}^{H\times W\times 1}$ is the gating weight for the l-th layer that is obtained from input features via channel split operation, $\odot$ is the element-wise multiplication operation, and M^i_{cs} is the aggregated features in the i-th branch. We denote the modulator in the last branch as M^b_{cs}. Since the modulator is adjusted according to the cross-scale intra-block features, the self-modulated manner can implicitly and efficiently improve feature representation. Finally, we adopt the element-wise multiplication to modulate the query features Q_{cs}. Therefore, the output $\hat{F}_{cs}$ can be computed as follows:

$$\hat{F}_{cs} = f^{prj2}_{cs}(Q_{cs} \odot f^{prj1}_{cs}(M^b_{cs})), \tag{4}$$

where $f^{prj1}_{cs}(\cdot)$ and $f^{prj2}_{cs}(\cdot)$ are the 1×1 convolutional layers, which enable the modulated features to fully interact across various channels.

2) Frequency Modulator: As shown in Fig. 4 (c), the FM contains two branches. Each branch learns respective feature representations within the Fourier domain. According to the theory of Fourier transform and previous work [16], [25], **(i)** the Fourier transform in a 2D image can be decomposed into the x-axis and y-axis, representing diffraction in the horizontal and vertical directions, respectively; **(ii)** processing information in the Fourier domain using 1×1 convolution can reflect the global features of the image; **(iii)** the amplitude spectrum mainly reflects the global haze-related information, and phase represents the texture structure for image dehazing. Therefore, we perform convolutional operations on different spectrums at two stages. *Subsequently, we will take the amplitude spectrum (the first stage) as examples to illustrate.*

In the first branch, we perform frequency learning in whole channel-dependent feature maps. Given an input feature maps F_{fre}, we firstly utilize the fast Fourier transform (FFT) [55] to obtain frequency features:

$$F^{real}_{fre}, F^{imag}_{fre} = \mathcal{FFT}(F_{fre}), \tag{5}$$

where F^{real}_{fre} and F^{imag}_{fre} represent the real and imaginary parts of the spectrums. The amplitude component A_{fre} and phase component P_{fre} can be obtained based on real and imaginary features.

$$A_{fre} = \left[{F^{real}_{fre}}^2(u,v) + {F^{imag}_{fre}}^2(u,v) \right]^{1/2},$$
$$P_{fre} = \arctan\left[\frac{F^{real}_{fre}(u,v)}{F^{imag}_{fre}(u,v)} \right], \tag{6}$$

where u and v denote the horizontal and vertical coordinates. Then, two cascaded 1×1 convolutions followed by a GELU [56] activation function are performed on the amplitude information, and we denote the produced new amplitude features as A'_{fre}. Next, we further convert the processed new Fourier features A'_{fre} and P_{fre} to their original space by using the inverse FFT operation. Considering that each channel of the resulting features differs in structure and haze degradation, we perform the channel-separated representational learning using a 3×3 depth-wise convolution. The process can be denoted as:

$$F'_{fre} = \mathrm{DConv}_{3\times 3}(\mathcal{FFT}^{-1}(A'_{fre}, P_{fre})), \tag{7}$$

where the F'_{fre} is the output of the first branch.

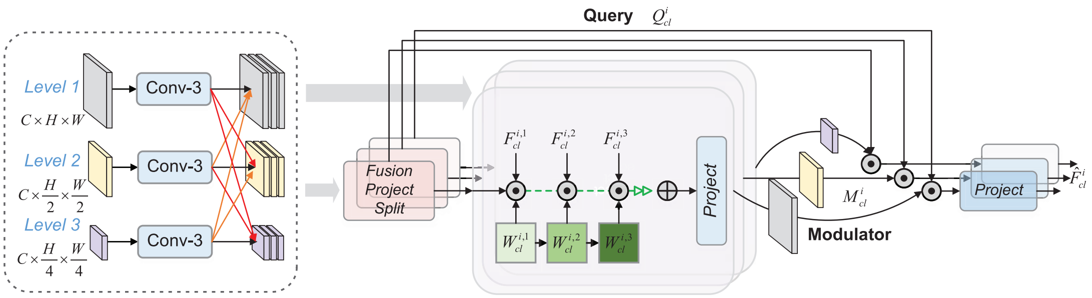


Fig. 5. The detailed illustration of the proposed cross-level modulator (CLM), which adopts a mutual-modulated manner to integrate inter-block features.

The second branch is designed to enhance the reconstruction of areas containing severe degradation or complex structures. Drawing inspiration from CBAM [57], our approach begins by compressing the input feature map along the channel dimension using both max pooling and mean pooling operations. Subsequently, the general content feature can be obtained via a 3 × 3 convolution,

$$F_g = \mathrm{Conv}([\mathrm{GAP}(F_{fre}), \mathrm{GMP}(F_{fre})]), \tag{8}$$

where $\mathrm{GAP}(\cdot)$ and $\mathrm{GMP}(\cdot)$ represent the global mean pooling and global max pooling respectively, $[\cdot,\cdot]$ denotes the channel concatenation operation, and $\mathrm{Conv}(\cdot)$ denotes convolutional operation. The obtained $F_g \in \mathbb{R}^{H\times W\times 1}$ contains important locations to focus on, which denotes channel-independent global representation. Moving forward, similar to the first branch, we also transform the F_g into amplitude and phase spectrum features in the Fourier domain, and then conduct the corresponding spectrum feature learning. Finally, we convert the frequency features to format the new spatial features F''_{fre} by adopting the inverse FFT.

To integrate features from dual branches, we further empower the single-channel content features F''_{fre} to modulate the channel-dependent global features F'_{fre} using the broadcasted calculation. This process is expressed as follows:

$$\hat{F}_{fre} = F'_{fre} \odot F''_{fre}, \tag{9}$$

where $\hat{F}_{fre}$ is the final output of this module. Compared with the previous Fourier learning schema, our FM selectively modulates essential information, thus improving haze removal and reconstructing image structure more accurately.

C. Cross-Level Modulator (CLM)

The SFM employed in each level empowers the network to capture intra-block contextual information. However, effectively utilizing cross-level information in both the encoder and decoder is crucial for enhancing inter-block feature modulation. Simple channel concatenation operations, as used in each level of the encoder-decoder, may not effectively leverage cross-level complementary information. To overcome this issue, we design a CLM to interconnect the deep features from their respective scales, as shown in Fig. 5.

The CLM is designed based on the idea of the CSM. However, there are several different points to emphasize. **(i)** For the multi-level cross-scale features, we first utilize the downsample and upsample operations to enable both top-down and bottom-up information flow. This procedure combines features from different receptive fields and enriches contextual representations. *Taking the i-th level of the encoder or decoder as instance*, we denote the enhanced cross-scale featues as $\{F_{cl}^{i,l}\}_{l=1}^{3}$, which corresponds to the $\{F_{cs}^{i,l}\}_{l=1}^{3}$ in the CSM. **(ii)** To formulate the query features and gating weights, we first adopt channel concatenation and 1 × 1 convolution to fuse the enhanced cross-scale features, and then adopt the channel Split operation to respectively obtain Q_{cl}^{i} and W_{cl}^{i}. **(iii)** Since the contextual feature dependencies have been captured based on the intra-block manner, we do not adopt additional average pooling operations similar to the CSM. Therefore, the cross-level aggregation can be written as:

$$Z_{cl}^{i,out} = \sum_{\ell=1}^{3} W_{cl}^{i,l} \odot F_{cl}^{i,l}, \tag{10}$$

where $Z_{cl}^{i,out}$ denotes the cross-level aggregated features in the i-th level. Further, the cross-level modulator M_{cl}^{i} can be attained by performing a feature projection convolution. Therefore, the final outputs in the i-th level are given as:

$$\hat{F}_{cl}^{i} = f_{cl}^{prj2}(Q_{cl}^{i} \odot f_{cl}^{prj1}(M_{cl}^{i})), \tag{11}$$

where $f_{cl}^{prj1}(\cdot)$ and $f_{cl}^{prj2}(\cdot)$ denote 1×1 convolution, $\hat{F}_{cl}^{i}$ denotes the ouput of the i-th level. The outputs at all three levels are adaptively modulated by cross-level features obtained through SFMs, facilitating the mutual modulation of inter-block features.

D. Loss Function

Our method adopts a two-stage architecture to construct the overall network. Therefore, the loss function consists of two parts: $\mathcal{L}_{stage1}$ and $\mathcal{L}_{stage2}$. Moreover, in the frequency domain, both stages operate on amplitude spectra and phase spectra, respectively. It is crucial to provide the corresponding supervision for these processes. Let I_{gt} denote the clean image. Then, the loss function can be denoted as:

$$\begin{aligned} \mathcal{L}_{stage1} &= \left\| O_1 - I_{gt} \right\|_1 + \alpha \left\| \mathcal{A}(O_1) - \mathcal{A}\left(I_{gt}\right) \right\|_1, \\ \mathcal{L}_{stage2} &= \left\| O_2 - I_{gt} \right\|_1 + \beta \left\| \mathcal{P}(O_2) - \mathcal{P}\left(I_{gt}\right) \right\|_1, \end{aligned} \tag{12}$$

where the first term and the second term in each equation above are performed on spatial and frequency domains, respectively, α and β are the trade-off factors, we empirically set them

TABLE I

Quantitative Comparison of State-of-the-Art Methods for image dehazing. The Best Results are bold

Method	SOTS-Indoor		SOTS-Outdoor		O-HAZE		Dense-Haze		NH-HAZE		#Params	FLOPs	Time
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	(M)	(G)	(S)
(TPAMI'10) DCP [30]	16.61	0.855	19.14	0.861	16.78	0.653	10.06	0.385	10.57	0.520	-	-	-
(TIP'16) DehazeNet [11]	19.82	0.821	24.75	0.927	17.57	0.770	13.84	0.425	16.62	0.524	0.009	0.581	1.708
(ICCV'17) AODNet [6]	20.51	0.816	24.14	0.920	15.03	0.540	13.14	0.410	15.40	0.570	0.002	0.115	0.217
(AAAI'20) FFA-Net [14]	36.59	0.989	33.57	0.984	22.12	0.770	14.39	0.452	19.87	0.692	4.456	287.5	2.345
(CVPR'20) MSBDN [35]	32.77	0.981	34.81	0.986	24.36	0.741	15.37	0.486	19.23	0.706	31.35	41.54	0.591
(CVPR'21) AECR-Net [9]	37.17	0.990	-	-	-	-	14.88	0.505	19.92	0.672	2.611	43.04	0.258
(CVPR'21) Restormer [18]	38.88	0.991	-	-	23.58	0.768	15.78	0.548	-	-	26.10	141.0	1.349
(CVPR'22) DeHamer [17]	36.63	0.988	35.18	0.986	25.11	0.777	16.62	0.560	20.66	0.684	132.4	59.67	0.585
(TIP'22) SGID-PFF [58]	38.52	0.991	30.20	0.975	20.96	0.741	12.49	0.517	-	-	18.90	81.13	0.731
(ECCV'22) FSDGN [16]	38.63	0.990	-	-	-	-	16.91	0.581	19.99	**0.731**	2.731	19.59	0.501
(ICCV'23) MBTFormer-B [20]	40.71	0.992	37.42	0.989	25.05	0.788	16.66	0.560	-	-	2.662	38.50	2.875
(ACM MM'23) MITNet [28]	40.23	0.992	35.18	0.988	-	-	16.97	0.606	21.26	0.712	2.731	16.42	0.584
(TIP'24) DehazeFormer-M [19]	38.46	0.994	34.29	0.983	-	-	16.29	0.510	20.47	**0.731**	4.634	48.64	1.123
(TIP'24) DEA-Net [42]	40.20	0.993	36.03	0.989	25.81	0.779	16.82	0.599	20.71	0.702	3.653	32.23	0.589
(TCSVT'24) OKNet [59]	40.79	0.993	37.68	0.989	25.64	0.784	16.92	0.608	20.48	0.712	4.720	39.67	0.612
(Ours) SFMN	**41.44**	**0.995**	**37.72**	**0.991**	**26.06**	**0.793**	**17.91**	**0.632**	**21.27**	0.707	3.424	18.39	0.573

as 0.05 according to MITNet [28], $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ denote the amplitude and phase components, respectively. Thus, the total loss $\mathcal{L}_{total}$ is denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{stage1} + \mathcal{L}_{stage2}. \tag{13}$$

IV. Experimental Results

A. Experimental Setup

Following recent works [17], [20], we employ the Indoor Training Subset (ITS) and Outdoor Training Subset (OTS) of the RESIDE dataset [21] as natural daytime image dehazing training data, and evaluate the resulting models on the corresponding test sets, *i.e.* SOTS-Indoor and SOTS-Outdoor, respectively. To assess the robustness of our SFMN in real-world scenarios, we utilize three commonly used real-world images, such as O-HAZE [60], Dense-Haze [61], and NH-HAZE [62]. Further, image desnowing datasets, including CSD [63], SRRS [64], and Snow100K [65], are adopted to verify the generalization abilities of our method. In addition, Peak Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index Measurement (SSIM) [66] are utilized for evaluation, and the metric code is based on the FFA-Net [14] and MSBDN [35]. We also report the model size and FLOPs, where the latter is calculated on the $3 \times 256 \times 256$ image patch. To further demonstrate the efficiency of various models, we choose the Dense-Haze dataset with a resolution of 1600×1200 to evaluate the average inference time.

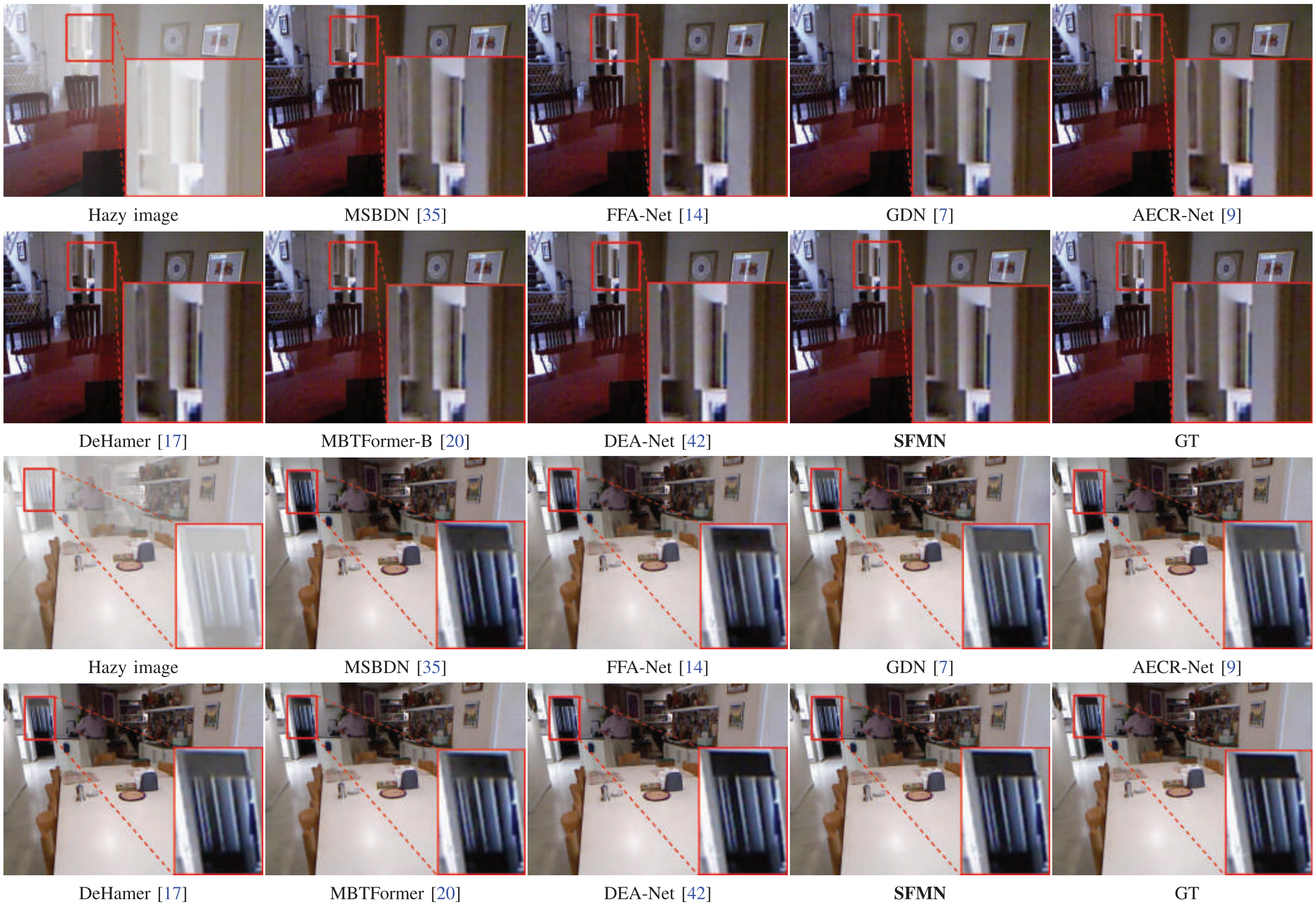
The number of SFM in each layer is set as $N = 1$. In each stage, each level of encoder and decoder from high-to-low resolution features is equipped with 20, 40, 80, and 160 channels, respectively. Besides, we use 2×2 transposed convolution and 4×4 strided convolution as the upsampling and downsampling layers. However, the upsampling and downsampling operation in the CLM is followed by previous work [53] for reducing model size. During training, we employ the ADAM optimizer [67] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Within each mini-batch for different datasets, we augment training samples by applying horizontal or vertical flips and rotations with 90°, 180°, and 270°. Our SFMN is implemented with the PyTorch [68] on an NVIDIA Tesla V100 GPU. Unless specifically emphasized, the patch and batch sizes are set to 256×256 and 16, respectively.

Our models are trained for 800 epochs on the ITS dataset and 40 epochs on the OTS dataset. As for real-world dehazing datasets, we train all models for a total of 4000 epochs on an 800×800 patch size, and the batch size is set to 4. As for image desnowing datasets, our models are trained for 1000K iterations. The initial learning rate of all models mentioned above is set to 2×10^{-4} and gradually reduced to 2×10^{-6} with the cosine annealing.

B. Experiments on Synthesized Dehazing Datasets

We compare SFMN with 14 CNN- and Transformer-based image dehazing methods, including DCP [30], DehazeNet [11], AODNet [6], FFA-Net [14], MSBDN [35], AECR-Net [9], Restormer [18], DeHamer [17], SGID-PFF [69], FSDGN [16], MITNet [28], MBTFormer-B [20], DehazeFormer-M [19], DEA-Net [42], and OKNet [59]. For those methods that did not provide pre-trained models, we re-trained them according to the provided codes.

1) Quantitative Evaluation: Table I shows the quantitative results on five datasets, from which we observe that the proposed SFMN attains 41.44 dB and 37.72 dB PSNR values on the SOTS-Indoor and SOTS-Outdoor datasets, respectively. Compared to the SOTA method MBTFormer-B [20], our approach achieves 0.73 dB and 0.30 dB PSNR performance gains, respectively. Additionally, compared with the recent Fourier transform-based method FSDGN [16], SFMN improves PSNR by 2.81 dB while maintaining a lower computational cost (FLOPs: 19.59 *vs.* 18.39). This indicates that our method excels at leveraging the spatial and frequency global features to improve performance. Compared to the recent OKNet, our method has obvious advantages on all datasets. Furthermore, we also evaluate the performance on real-world datasets. As we can see, our SFMN performs better than all previous methods, except for the SSIM value on the NH-


Hazy image MSBDN [35] FFA-Net [14] GDN [7] AECR-Net [9]

DeHamer [17] MBTFormer-B [20] DEA-Net [42] **SFMN** GT

Hazy image MSBDN [35] FFA-Net [14] GDN [7] AECR-Net [9]

DeHamer [17] MBTFormer [20] DEA-Net [42] **SFMN** GT

Fig. 6. Visual comparisons on the SOTS [21] dataset. Zoom in for a better view.

TABLE II

COMPARISON OF NIQE AND FADE FOR DEHAZING METHODS IN THE LAST THREE YEARS. THE BEST RESULTS ARE BOLD

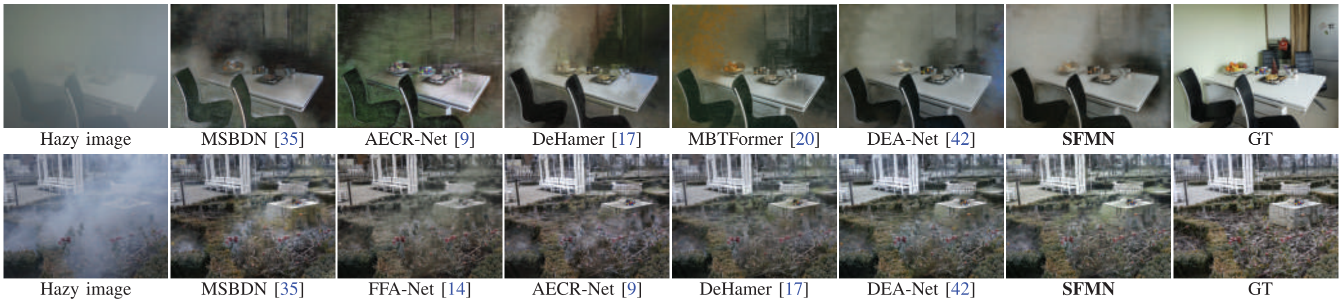
Method	SOTS-Indoor		SOTS-Outdoor	
	NIQE ↓	FADE ↓	NIQE ↓	FADE ↓
DeHamer [17]	4.648	0.5020	3.009	0.8318
SGID-PFF [58]	4.892	0.5219	3.428	0.8891
MITNet [28]	4.253	0.5019	2.945	0.7562
MBTFormer-B [20]	4.421	0.4992	2.998	0.8472
DehazeFormer [19]	4.328	0.4987	2.951	0.8093
DEA-Net [42]	4.381	0.4965	2.963	0.7592
OKNet [59]	4.267	0.4961	2.961	0.7651
(Ours) SFMN	**4.221**	**0.4953**	**2.944**	**0.7452**

HAZE. What is even more remarkable is that our approach obtains a better balance between model complexity and performance compared to Transformer-based methods such as DeHamer, Restormer, MBTFormer-B, and DehazeFormer-M. These findings substantiate the effectiveness and efficiency of our contextual feature modeling. In Table II, we further evaluate NIQE and FADE of recent three-year dehazing methods on the SOTS-Indoor and SOTS-Outdoor datasets. It can be seen that SFMN achieved the lowest results on both datasets, indicating that its perception effect and dehazing ability are better.

2) *Visual Comparisons:* In Fig. 6, we can observe that other competitors successfully remove most of the haze but appear to have colorfulness distortion or texture loss. In contrast, our method effectively removes the homogeneous haze and reconstructs vivid texture and details. Fig. 7 presents two samples from the Dense-Haze and NH-HAZE datasets. As we know, these two datasets are generated by professional haze machines and thus contain non-homogeneous and thicker haze than the RESIDE dataset. We found that all methods have difficulty in restoring results close to the ground-truth image and produce color distortions. In comparison, our method produces slight artifacts and less residual haze, resulting in a better overall appearance.

C. *Experiments on Real-World Dehazing Datasets*

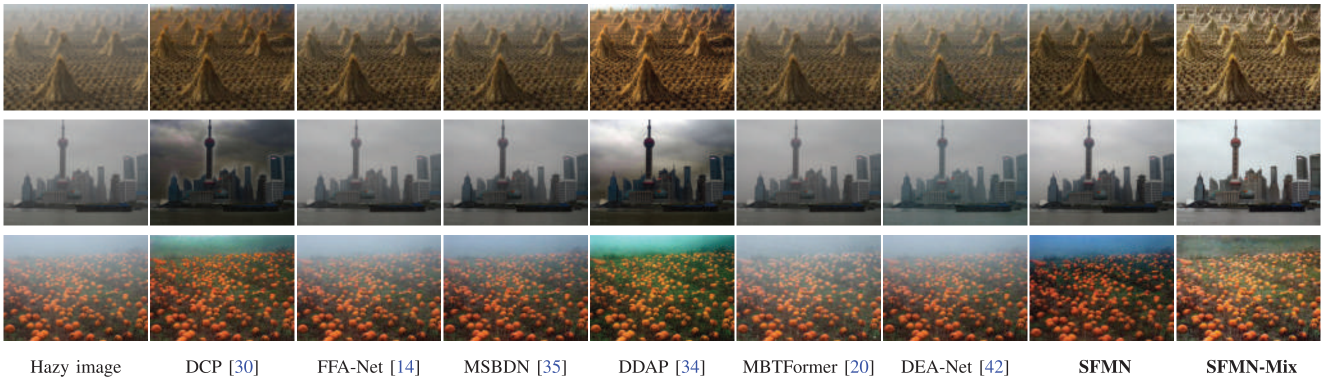
We evaluate the performance of the proposed SFMN on the natural real-world hazy images, whose corresponding haze-free images are not available. The quality index NIQE [75] and FADE [76] are used to evaluate 29 real-world dehazing images [34], and the results are shown in Table III. DCP [30] and DDAP [34] are two prior-based algorithms, and they can perform better than data-driven-based methods. Due to the domain gap between synthetic and real-world data, our method and other CNN-based methods indeed perform worse than prior-based methods. However, as observed, our SFMN achieves better than purely data-driven methods. To achieve



Hazy image MSBDN [35] AECR-Net [9] DeHamer [17] MBTFormer [20] DEA-Net [42] **SFMN** GT

Hazy image MSBDN [35] FFA-Net [14] AECR-Net [9] DeHamer [17] DEA-Net [42] **SFMN** GT

Fig. 7. Visual comparisons on Dense-Haze [61] and NH-HAZE [62] datasets. Zoom in for a better view.



Hazy image DCP [30] FFA-Net [14] MSBDN [35] DDAP [34] MBTFormer [20] DEA-Net [42] **SFMN** **SFMN-Mix**

Fig. 8. Visual results of various methods on real-world hazy images, where haze-free images are not available. Zoom in for a better view.

TABLE III

Comparison of Quantitative Results on Real-World Rainy Images, and Note That Lower Scores Indicate Better Image Quality

Methods	Hazy input	DCP [30]	FFA-Net [14]	MSBDN [35]	DDAP [34]	MBTFormer [20]	DEA-Net [42]	**SFMN**	**SFMN-Mix**
NIQE↓/FADE↓	3.94/3.0055	**3.27**/0.4579	4.05/2.7930	3.77/1.9847	3.28/**0.4237**	3.84/1.8884	3.72/2.0573	3.61/1.7381	3.35/0.8785

TABLE IV

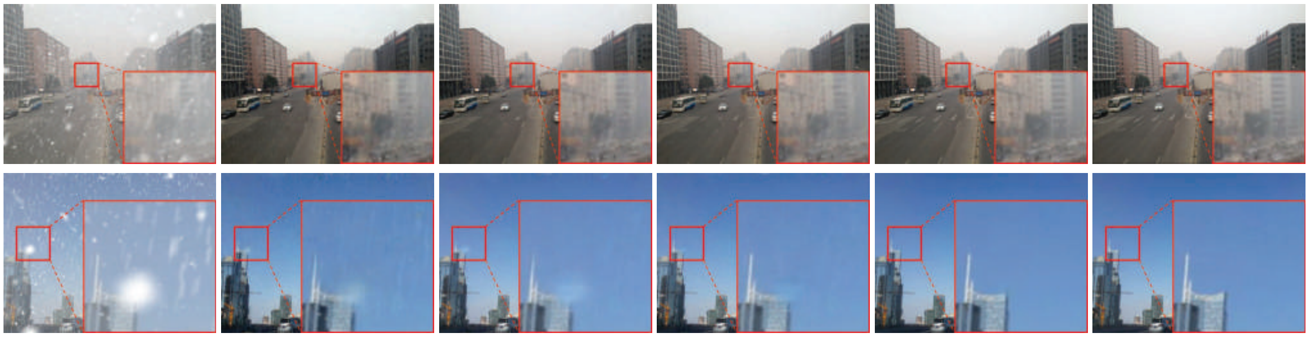
Quantitative Comparison for Image Desnowing on CSD [63], SRRS [64], and Snow100K [65] Datasets. The Best Results are Bold

Dataset	Metric	DesnowNet [65]	CycleGAN [70]	All in one [71]	JSTASR [64]	HDCW-Net [63]	TransWeather [72]	NAFNet [73]	Restormer [18]	MSPFormer [74]	**SFMN** **(Ours)**
CSD	PSNR	20.13	20.98	26.31	27.96	29.06	31.76	33.13	35.45	33.57	**37.50**
	SSIM	0.81	0.80	0.87	0.88	0.91	0.93	0.96	0.97	0.96	**0.98**
SRRS	PSNR	20.38	20.21	24.98	25.82	27.78	28.29	29.72	32.24	30.76	**35.21**
	SSIM	0.84	0.74	0.88	0.89	0.92	0.92	0.94	0.96	0.95	**0.98**
Snow100K	PSNR	30.50	26.81	26.07	23.12	31.54	31.82	32.41	**34.67**	33.43	34.43
	SSIM	0.94	0.89	0.88	0.86	0.95	0.93	0.95	0.95	**0.96**	0.94

better real-world image removal, we follow DehazeFormer [19] and train the network using a synthetic OTS, Dense-HAZE, and NH-HAZE mixed dataset to obtain SFMN-Mix. The results show that the model can achieve competitive performance, similar to the prior-based method. As shown in Fig. 8, the prior-based method DCP [30] and DDAP [34] perform more photo-realistic than other data-driven methods, including FFA-Net [14], MSBDN [35], DEA-Net [42] and MBTFormer [20], while some slight artifacts can also be produced, especially in the sky regions. In contrast, our method produces balanced visual dehazing results. It is particularly noteworthy that the model SFMN-Mix trained on the mixed dataset obtained very visually pleasing results. This is partly because our SFMN has a powerful discriminative representation ability that incorporates spatial and frequency contexts. All 29 real-world compared results can be found in the supplementary material.

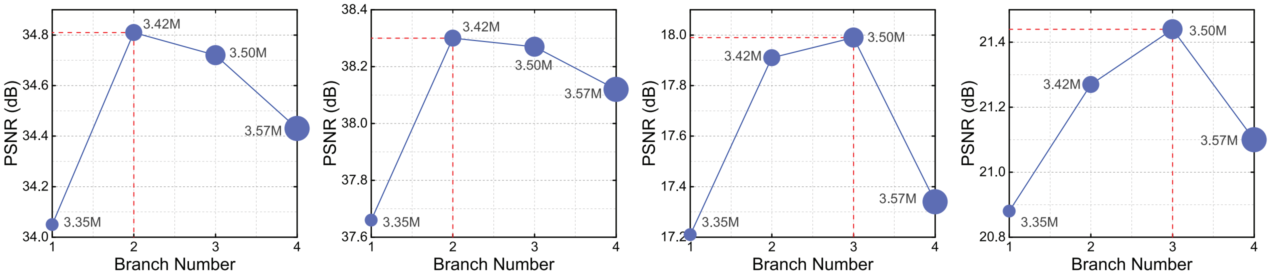
D. Experiments on Image Desnowing

To evaluate the generalization capabilities of our SFMN in other low-level tasks, we further conduct experiments on image desnowing datasets (*e.g.,* CSD [63], SRRS [64], and Snow100K [65]), showing the quantitative results in the Table IV. Notably, our approach can lead to most image desnowing methods on three datasets. Among them, Restormer [18] and NAFNet [73] are universal image restoration models that employ spatial global modeling mechanisms. However,



Input image DesnowNet [65] JSTASR [64] HDCW-Net [63] **SFMN** GT

Fig. 9. Visual comparisons of image desnowing methods on the CSD [63] dataset. Zoom in for a better view.



(a) ITS [21] (b) OTS [21] (c) Dense-Haze [61] (d) NH-HAZE [62]

Fig. 10. The ablation study of the number of branches in CSM on different image dehazing datasets.

our method only performs slightly worse on one dataset, highlighting that the schema of incorporating cross-scale, cross-level, and frequency feature modulation together is effective for image desnowing tasks. Fig. 9 shows desnowing visual comparisons on the CSD [63] dataset, from which we observe the proposed SFMN can output more visually pleasing desnowing results, however, other competitors still retain slight degradation and produce blurry reconstructed images.

E. Model Complexity Analysis

To comprehensively showcase the efficiency and effectiveness of the proposed SFMN, we further compare the number of parameters, number of FLOPs, and average inference time with the SOTA methods over the last three years. As shown in Fig. 2 and Table I, our method achieves the best performance with the lowest FLOPs, relatively low parameters, and relatively faster inference speed. FSDGN [16] and AECR-Net [9] utilize fewer parameters to achieve faster inference speed, but there is a large performance gap compared with our method. MBTFormer-B [20] and DeHamer [17] adopt the Transformer architecture, thereby consuming more time to deal with images. In contrast, we jointly adopt spatial-frequency design to mine feature dependency relationships, reaching a good trade-off between model complexity and performance. In addition, we also evaluate the inference time on high-resolution images and observe that the Transformer-based MBTFormer-B has an extremely slow inference speed, which is mainly caused by the self-attention operation.

F. Ablation Study

We investigate the effectiveness of our proposed modules and architecture design. For all ablation studies, we train our models on the ITS [21] dataset for 200K iterations with the initial learning rate as $2e^{-4}$ and batch size as 16. Unless specified in the table, other configurations are identical to that of our final dehazing model.

1) The Ablation Study of CSM: We first conduct ablation studies to clarify the influence of branch numbers in the CSM on four datasets. As shown in Fig. 10, the vertical axis is PSNR, and the circle radius is the model parameters. As can be observed, the performance gains on different datasets are different. For the low-resolution ITS and OTS datasets, employing two branches can achieve the best performance. When equipped with more branches, the performance of the corresponding models decreases, possibly because the disadvantage of reducing the size of low-resolution features, resulting in the loss of too much spatial information, outweighs the advantages of multi-scale learning. Nonetheless, the performance maximums are achieved when employing three branches for the high-resolution Dense-Haze and NH-HAZE datasets. The main reason is that these two datasets include higher-resolution images, and when the images undergo large-scale downsampling, the finer details are lost. Since we aim to pursue efficient image dehazing, we uniformly set two branches for all datasets. Fig. 11 shows the feature variation when the branch number is three, and we observe that the obtained feature exhibits a coarse-to-fine form. This is because low-scale features have a larger receptive field, which is also consistent with our motivation. We also study the impact of

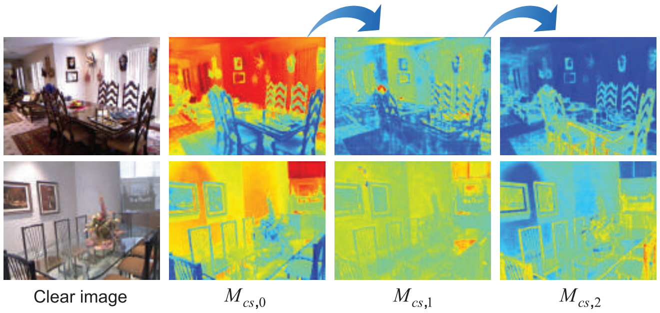


Fig. 11. Visualization of features in CSM when the branch number is set to three, where $M_{cs,i}$ denotes the output of i-th branch in the CSM.

TABLE V
ABLATION STUDY OF KERNEL SIZE IN THE CSM

Setting	All k=3	All k=5	All k=7	**Ours**	k=[5,7,9]	k=[7,9,11]
Params (M)	3.33	3.41	3.54	3.42	3.55	3.71
FLOPs (G)	17.72	18.29	19.15	18.39	19.24	20.38
PSNR (dB)	37.80	38.13	38.37	38.30	**38.41**	38.24

TABLE VI
ABLATION STUDY OF THE FM

Setting	#Params (M)	FLOPs (G)	PSNR (dB)
w/ Branch 1	3.42	18.37	38.04
w/ Branch 2	3.25	18.01	37.88
Ours (Element-wise Multi.)	3.42	18.39	**38.30**
Element-wise Sum	3.42	18.39	38.17

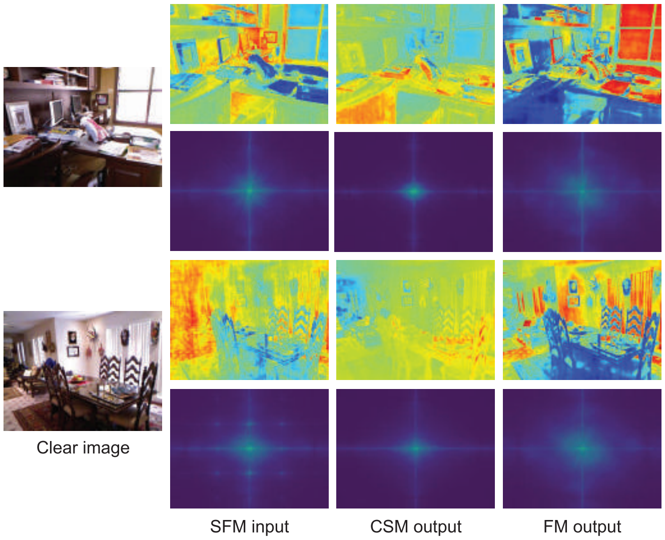


Fig. 12. Visualization of the feature map and Fourier spectral features before and after using the proposed components, including CSM and FM.

TABLE VII
BREAK-DOWN ABLATION STUDY FOR BETTER EVALUATION OF THE PROPOSED COMPONENTS, INCLUDING CSM AND FM

Model	PSNR (dB)	#Params (M)	FLOPs (G)
(a) Baseline-Conv	35.21	2.92	14.07
(b) Baseline-CSM	37.25	2.65	13.26
(c) (a) + FM	36.28	3.26	15.18
(d) (b) + FM	37.99	2.99	14.39
(e) (d) + CLM	38.30	3.42	18.39
(f) (d) + CLM-OP	38.07	3.42	18.39

kernel size in CSM for performance in Table V. Two groups of experiments are conducted: one involves keeping all kernel sizes the same, while the other progressively increases the kernel size at each level. As can be observed: **(i)** opting for a moderate kernel size (All $k = 7$) yields slightly better performance than choosing a small kernel size; **(ii)** employing too large a kernel size at each level (k = [7,9,11]) may lead to performance degradation, this could be attributed to the limitations of small kernel sizes in capturing long-range dependencies and the loss of local structural information with overly large kernel sizes. However, our setting can progressively fuse local-to-global features. When k = [5,7,9] or $k = 7$, the result performs better than ours yet comes at the cost of longer training times and increased expenses. For a better performance/cost trade-off, we set k = [3,5,7] in the final models.

2) The Ablation Study of FM: We alternately utilize the Fourier transform in two branches of the FM to conduct validations. The first branch focuses on learning from the channel-dependent global features, while the second branch emphasizes learning from the channel-independent single content features based on crucial regions. Thus, utilizing the latter to modulate the former can achieve effective feature selection. In Table VI, the results after individually using them perform inferiorly to the final model, consistent with our analysis. Besides, the effect of element-wise sum for fusion is not as good as element-wise multiplication, which further proves the rationality of our design.

3) Component Analysis: **Firstly**, we perform a breakdown ablation to explore the effect of each component and their interaction. The baseline network (a) is derived by adopting three cascaded 3 × 3 convolutions to replace SFM and removing the CLM from SFMN. As indicated in Table VII, the model attains 35.21 dB PSNR, underscoring the superiority of our two-stage design. Then, we embed the CSM into the network to construct the other baseline, named model (b), and the model achieves 2.04 dB performance gains. Based on models (a) and (b), we further embed the FM into the corresponding networks, and the corresponding models (c) and (d) achieve additional performance boosts of 1.07 dB and 0.74 dB, respectively, indicating that the frequency modulation mechanism can cooperate with our spatial contextual modulation mechanism well. Finally, when we merge CLM into model (d), model (e) can obtain the best performance, showing its significant contribution. To eliminate the doubt that the performance gain is due to the increase in parameters, we add a new ablation experiment, that is, directly adding the relevant convolution operations in CLM to the network to ensure that the number of parameters and FLOPs of the model remain unchanged, recorded as model (f). As can be seen from the table, the performance of the model has only increased by 0.08 dB. The specific reason is that the model has only undergone some separable depth-wise convolutions for feature extraction, but lacks further interaction

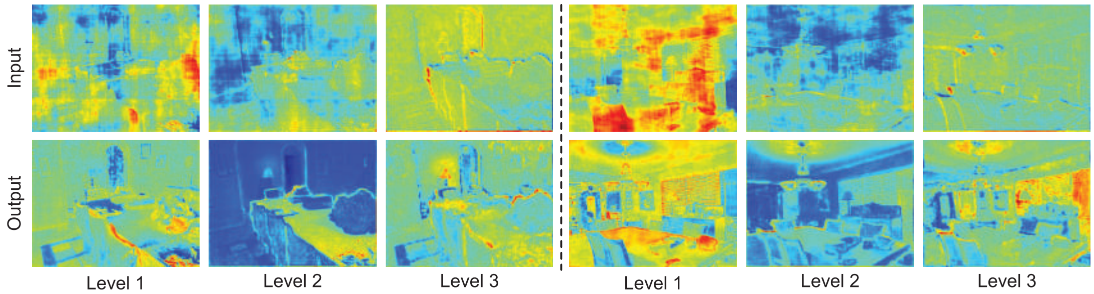


Fig. 13. Visualization of the feature map before and after using CLM. Here, we present all results in the encoder of the first-stage network.

TABLE VIII

FURTHER STUDIES OF CSM AND CLM, INCLUDING THE MODULATION AND GATING MECHANISMS

Model	PSNR (dB)	#Params (M)	FLOPs (G)
CSM w/o gating	38.09	3.42	18.39
CSM w/o modulation	37.88	3.42	18.39
CLM w/o gating	38.16	3.42	18.39
CLM w/o modulation	38.11	3.42	18.39
Ours	38.30	3.42	18.39

between features. **Secondly**, we microscopically investigate the influence of the gating mechanism in Eqs. (3) and (10) and modulation mechanism (element-wise multiplication) in Eqs.(4) and (11) for performance. As shown in Table VIII, removing these operations leads to an obvious performance drop yet maintains the same parameters and FLOPs. These show that **(i)** the gating mechanism helps the model to learn to aggregate useful information adaptively, and **(ii)** adopting element-wise multiplication to modulate initial features is a simple and powerful way, which further enhances feature second-order interaction.

Fig. 12 visualizes the output features from the proposed SFM. We found that spatial features and frequency features focus on different aspects. In detail, the CSM not only highlights the edge information but also shows clearer outlines. In addition, FM operates in the Fourier domain and focuses on both high and low frequency information. Therefore, the final produced features contain abundant and rich global and local contextual information, which is vital to reconstructing image structure information.

We visualize the feature maps before and after applying the CLM shown in Fig. 13. As we all know, high-resolution features usually lack awareness of contextual information. However, they contain abundant geometric information, such as points, edges, textures, and so on. In contrast, the low-resolution features are short in geometric information but contain adequate contextual information. To this end, our CLM can better exploit the within-scale characteristics and the cross-level complementarity. After using the module, we discover that **(i)** high-resolution features (Level 1) not only contain more fine-grained details but also add rich contextual information, and **(ii)** low-resolution features (Level 3) replenish more details and explicit textures, further verifying the necessity of cross-level hierarchical feature modeling.

TABLE IX

THE ABLATION STUDY OF THE DESIGN OF SFM

Model	(a) FM → CSM	(b) CSM $\|$ FM	(c) CSM→FM (ours)
PSNR (dB)	37.18	38.06	38.30

TABLE X

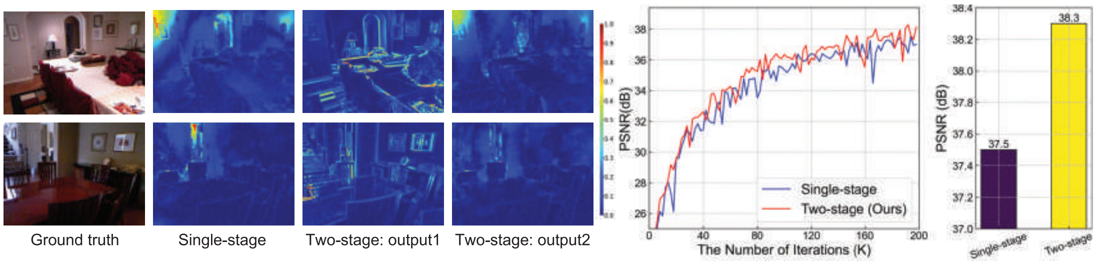
THE QUANTITATIVE COMPARISONS BEFORE AND AFTER EMPLOYING THE CLM TO RESTORMER [18] AND NAFNET [73]

Model	Restormer	Restormer w/ CLM	NAFNet	NAFNet w/ CLM
PSNR (dB)	36.01	36.27 (+0.26)	37.13	37.50 (+0.37)
Parmas (M)	3.02	3.12	29.10	30.77
FLOPs (G)	17.28	18.16	16.23	21.38

4) The Design of Spatial Frequency Modulator (SFM): In the final model, we stack the CSM and FM modules to construct the SFM sequentially. Here, we explore more ways to build this module. As shown in Table IX, model (a) (denoted as FM → CSM) and model (c) (denoted as CSM → FM) adopt sequential manner while having a different order of CSM and FM, model (b) (CSM || FM) adopts the parallel format to design SFM. As we can see, model (a) presents the worst performance, and model (b) shows relatively low results. We consider that the frequency information can provide additional prior knowledge, while this requires the support of powerful spatial information. The paper demonstrates that using the frequency domain module alone does not yield remarkable results.

5) Extend Cross-Level Modulator (CLM) to Other Backbones: To further demonstrate the effectiveness of the proposed CLM, we conduct a group of experiments by plugging it into existing well-designed image restoration networks, such as Restormer [18] and NAFNet [73], to experiment on the SOTS-indoor dataset. As for Restormer, we set the number of Transformer blocks as $[2, 3, 3, 4]$ from level 1 to level 4, and the basic channel number is set to 16. The other configurations are consistent with the original paper. As for NAFNet, all settings are consistent with the original paper. As shown in Table X, the models obtain more performance gains after employing our CLM but with only a slight increase of FLOPs and parameters, which is mainly attributed to the superior cross-level contextual modeling capabilities of the CLM.

6) The Effectiveness of Two-Stage Design: We customize a single-stage network based on our proposed components,



(a) Error map comparisons.

(b) Convergence analysis.

Fig. 14. Ablation studies of two-stage design. (a) compares the error maps between dehazed images and clear images. (b) shows the training process for training single-stage and two-stage networks.

where two SFMs are embedded into each layer of the encoder and decoder to ensure similar model complexity and size. The comparisons of the error map and convergence curves are shown in Fig. 14. The error map between the dehazed image and the ground-truth image reflects the learning ability of the method in detail recovery [77]. Clearly, our two-stage design can preserve the structure and recover details more accurately, which can be observed from the smaller error produced. In addition, our method converges slightly faster than the single-stage method. These results fully demonstrate the rationality of our two-stage design.

V. Conclusion

This paper proposes a spatial frequency modulation network (SFMN) for image dehazing. The core design is efficient global modeling components from spatial and frequency domains. In detail, we design a basic embedding module, named spatial frequency modulator, based on the inter-block feature modulation manner. In the spatial domain, the cross-scale modulator is developed to gradually capture hierarchical contextual features to achieve spatial feature modulation. The frequency modulator is implemented in the frequency domain to achieve global haze removal and emphasize regions with severe haze degradation and complex structures. In addition, we construct a cross-level modulator to achieve inter-block feature mutual modulation in the encoder and decoder of the two-stage network. Extensive experiments demonstrate that our SFMN achieves superior performance over recent state-of-the-art methods. Meanwhile, extended experiments on image desnowing also prove the robustness of our method.

References

[1] Y. Liu et al., "Improved techniques for learning to dehaze and beyond: A collective study," 2018, *arXiv:1807.00202*.

[2] C. Sakaridis, D. Dai, S. Hecker, and L. V. Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 687–704.

[3] S. Zhou, D. Guo, J. Li, X. Yang, and M. Wang, "Exploring sparse spatial relation in graph inference for text-based VQA," *IEEE Trans. Image Process.*, vol. 32, pp. 5060–5074, 2023.

[4] D. Guo, K. Li, Z.-J. Zha, and M. Wang, "DADNet: Dilated-attention-deformable ConvNet for crowd counting," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1823–1832.

[5] 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.

[6] 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.*, Oct. 2017, pp. 4780–4788.

[7] 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.

[8] W. Ren et al., "Gated fusion network for single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3253–3261.

[9] H. Wu et al., "Contrastive learning for compact single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10551–10560.

[10] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5896–5905.

[11] 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.

[12] 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.

[13] 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.

[14] 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, vol. 34, no. 7, pp. 11908–11915.

[15] Y. Gou, P. Hu, J. Lv, J. T. Zhou, and X. Peng, "Multi-scale adaptive network for single image denoising," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Aug. 2022, pp. 14099–14112.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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 Int. Conf. Comput. Vis.*, Jun. 2023, pp. 12802–12813.

[21] B. Li et al., "Benchmarking single-image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2018.

[22] X. Ma, X. Dai, Y. Bai, Y. Wang, and Y. Fu, "Rewrite the stars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 5694–5703.

[23] B. Niu et al., "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 191–207.

[24] Y. Tay et al., "OmniNet: Omnidirectional representations from transformers," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2021, pp. 10193–10202.

[25] M. Zhou, J. Huang, C.-L. Guo, and C. Li, “Fourmer: An efficient global modeling paradigm for image restoration,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 42589–42601.

[26] 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.

[27] X. Cong, J. Gui, J. Zhang, J. Hou, and H. Shen, “A semi-supervised nighttime dehazing baseline with spatial-frequency aware and realistic brightness constraint,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2631–2640.

[28] 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.

[29] C. Li et al., “Embedding Fourier for ultra-high-definition low-light image enhancement,” 2023, *arXiv:2302.11831*.

[30] 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.

[31] 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.

[32] 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.

[33] Z. Li, J. Zheng, Z. Zhu, W. Yao, and S. Wu, “Weighted guided image filtering,” *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 120–129, Jan. 2014.

[34] Z. Li, H. Shu, and C. Zheng, “Multi-scale single image dehazing using Laplacian and Gaussian pyramids,” *IEEE Trans. Image Process.*, vol. 30, pp. 9270–9279, 2021.

[35] 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.

[36] 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.

[37] Z. Zheng and X. Jia, “4K-HAZE: A dehazing benchmark with 4K resolution hazy and haze-free images,” 2023, *arXiv:2303.15848*.

[38] P. Wang et al., “UHD image dehazing via anDehazeFormer with atmospheric-aware KV cache,” 2025, *arXiv:2505.14010*.

[39] Y. Zheng, J. Zhan, S. He, J. Dong, and Y. Du, “Curricular contrastive regularization for physics-aware single image dehazing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 5785–5794.

[40] Y. Zhang, S. Zhou, and H. Li, “Depth information assisted collaborative mutual promotion network for single image dehazing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2846–2855.

[41] 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.

[42] 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.

[43] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, “Model-based deep learning,” *Proc. IEEE*, vol. 111, no. 5, pp. 465–499, May 2023.

[44] S. Zhao, L. Zhang, Y. Shen, and Y. Zhou, “RefineDNet: A weakly supervised refinement framework for single image dehazing,” in *Proc. IEEE Trans. Image Process.*, vol. 30, Jan. 2021, pp. 3391–3404.

[45] Z. Li, C. Zheng, H. Shu, and S. Wu, “Dual-scale single image dehazing via neural augmentation,” *IEEE Trans. Image Process.*, vol. 31, pp. 6213–6223, 2022.

[46] W. Zou, M. Jiang, Y. Zhang, L. Chen, Z. Lu, and Y. Wu, “SDWNet: A straight dilated network with wavelet transformation for image deblurring,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 1895–1904.

[47] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, “Multi-level wavelet-CNN for image restoration,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 773–782.

[48] X. Song, D. Zhou, W. Li, H. Ding, Y. Dai, and L. Zhang, “WSAMF-net: Wavelet spatial attention-based MultiStream feedback network for single image dehazing,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 575–588, Feb. 2023.

[49] J. Huang et al., “Deep Fourier-based exposure correction network with spatial-frequency interaction,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 163–180.

[50] C. Wang, H. Wu, and Z. Jin, “FourLLIE: Boosting low-light image enhancement by Fourier frequency information,” in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7459–7469.

[51] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, “Intriguing findings of frequency selection for image deblurring,” 2021, *arXiv:2111.11745*.

[52] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[53] S. W. Zamir et al., “Multi-stage progressive image restoration,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, May 2021, pp. 14821–14831.

[54] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[55] M. Frigo and S. G. Johnson, “FFTW: An adaptive software architecture for the FFT,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Oct. 1998, pp. 1381–1384.

[56] D. Hendrycks and K. Gimpel, “Gaussian error linear units (GELUs),” 2016, *arXiv:1606.08415*.

[57] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[58] H. Bai, J. Pan, X. Xiang, and J. Tang, “Self-guided image dehazing using progressive feature fusion,” *IEEE Trans. Image Process.*, vol. 31, pp. 1217–1229, 2022.

[59] Y. Cui, W. Ren, and A. Knoll, “Omni-kernel modulation for universal image restoration,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 12, pp. 12496–12509, Dec. 2024.

[60] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, “O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 754–762.

[61] 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.

[62] 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.

[63] W.-T. Chen et al., “ALL snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4196–4205.

[64] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, “JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 754–770.

[65] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, “DesnowNet: Context-aware deep network for snow removal,” *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3064–3073, Jun. 2018.

[66] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[67] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[68] A. Paszke et al., “Automatic differentiation in Pytorch,” in *Proc. Conf. Neural Inf. Process. Syst. Workshops*, 2017, pp. 1–4.

[69] T. Ye et al., “Perceiving and modeling density is all you need for image dehazing,” 2021, *arXiv:2111.09733*.

[70] D. Engin, A. Genç, and H. K. Ekenel, “Cycle-dehaze: Enhanced cyclegan for single image dehazing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Oct. 2018, pp. 825–833.

[71] R. Li, R. T. Tan, and L.-F. Cheong, “All in one bad weather removal using architectural search,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3175–3185.

[72] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, “TransWeather: Transformer-based restoration of images degraded by adverse weather conditions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2343–2353.

[73] L. Chen, X. Chu, X. Zhang, and J. Sun, “Simple baselines for image restoration,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 17–33.

[74] S. Chen et al., “MSP-former: Multi-scale projection transformer for single image desnowing,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[75] 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.

[76] 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.

[77] C. Zheng, D. Shi, and Y. Liu, “Windowing decomposition convolutional neural network for image enhancement,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 424–432.

Hao Shen (Graduate Student Member, IEEE) received the Ph.D. degree from the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China, in 2024. He is currently a Lecturer with the School of Public Security and Emergency Management, Anhui University of Science and Technology, Hefei. He has published more than ten papers in conferences, such as AAAI, ACM MM, CVPR, ECAI, and ICME, as well as journals such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and PR. His research interests include computer vision and deep learning.

Henghui Ding (Member, IEEE) received the B.E. degree from Xi’an Jiaotong University, Xi’an, China, in 2016, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2020. He was a Research Scientist with ByteDance and a Postdoctoral Researcher with ETH Zurich and NTU. He is currently a Professor with Fudan University, Shanghai, China. His research interests include computer vision and machine learning. He serves as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP) and serves/served as the Area Chair of top conferences, such as CVPR, NeurIPS, ICML, ICLR, and ACM MM.

Yulun Zhang (Member, IEEE) received the B.E. degree from the School of Electronic Engineering, Xidian University, China, in 2013, the M.E. degree from the Department of Automation, Tsinghua University, China, in 2017, and the Ph.D. degree from the Department of ECE, Northeastern University, USA, in 2021. He was a Postdoctoral Researcher with the Computer Vision Laboratory, ETH Zurich, Switzerland. He is currently an Associate Professor with Shanghai Jiao Tong University, Shanghai, China. His research interests include image/video restoration and synthesis, biomedical image analysis, model compression, multimodal computing, large language models, and computational imaging. He is/was the Area Chair of CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, IJCAI, ACM MM, and a Senior Program Committee (SPC) Member for IJCAI and AAAI.

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 of image processing with the CNRS UMR6168, Laboratory Sciences de Information et des Systems, La Garde, France. From 2013 to 2014, he was a Research Fellow of 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 bachelor’s and master’s degrees in electrical and electronic engineering from the University of Electronic Science and Technology of China, and the Ph.D. degree in electrical and electronic engineering from Helmut Schmidt University, Hamburg, Germany. 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 Centre 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. He has authored over 200 articles with over 60 papers in IEEE journals and 30 papers in top conferences, such as CVPR/ICCV/ECCV/AAAI/NeurIPS/ICLR. His research interests include computer vision, machine learning, pattern recognition, image processing, and biometrics. He served as an Associate Editor for IEEE SIGNAL PROCESSING LETTERS and IEEE TRANSACTIONS ON IMAGE PROCESSING (IEEE TIP). He is also serving as a Senior Area Editor for IEEE TIP and the Editor-in-Chief for *IET Biometrics*.