Visual Intelligence

RESEARCH Open Access

Efficient image dehazing with synergic expert modulation

Hao Shen[1,2], Xiaofeng Cong[3], Henghui Ding[4*], Yulun Zhang[5] and Xudong Jiang[6]

Abstract

Recent developments in context modulation mechanisms have achieved significant improvements in performance, as well as better trade-offs between model accuracy and efficiency. These mechanisms operate on input through context modeling and then leverage these contexts to modulate projected input features. However, existing methods have limitations. First, they cannot adaptively learn the extracted hierarchical context or ignore the complementarity of the cross-scale context. Second, these methods do not adequately address the unique frequency characteristics of hazy images. In response, we propose a synergic expert modulation (SEM) mechanism to explicitly model context information. Specifically, the SEM consists primarily of two mixture of spatial experts (MSE) modules that handle features of different scales and one mixture of frequency experts (MFE) module that operates within the frequency domain. The MSE learns hierarchical features of various granularities in an adaptive manner, guided by multiple gating experts and a routing network. The MFE specializes in mining frequency contexts guided by multiple frequency experts and a routing network. At the micro level, each frequency expert operates in two stages: spectral filtering and spectral learning. The former performs mask filtering to enhance the weights of low-frequency components, and the latter performs Fourier amplitude and phase decoupled learning, thus promoting the removal of haze information and global context learning. Finally, the obtained contexts are integrated to modulate the projected feature, thereby significantly enhancing cross-domain feature synergies. The proposed network, referred to as the synergic expert modulation network, is constructed by inserting SEM-based building blocks into the U-Net architecture to increase efficiency. Extensive experiments demonstrate that our network achieves state-of-the-art performance on multiple datasets for the image dehazing task while incurring lower computational costs.

Keywords: Image dehazing, Mixture of spatial experts, Mixture of frequency experts

1 Introduction

Image clarity significantly impacts downstream visual tasks such as object detection [1, 2], scene understanding [3], and semantic segmentation [4]. Image dehazing, as a long-standing low-level visual task, aims to reconstruct haze-free high-quality images. Therefore, it has always received extensive attention from academic and industrial communities.

Existing image dehazing methods can be categorized into two main types: traditional approaches and deep learning-based approaches. The emergence of convolutional neural networks (CNNs) has greatly improved the performance of dehazing, which can implicitly or explicitly learn powerful image priors from large-scale data. The primary contribution of CNN-based methods is attributed to advanced architecture designs, including multi-scale [5], attention mechanism [6, 7], and contrastive learning [8].

*Correspondence: henghui.ding@gmail.com
[4]Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, 200433, China
Full list of author information is available at the end of the article

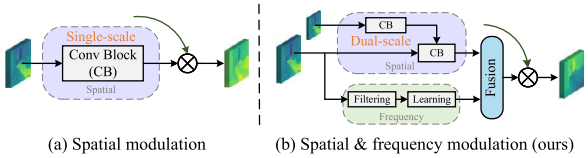

Figure 1 The illustration of modulation mechanisms. ⊗ is element-wise multiplication. Compared to solely using spatial modulation [13, 14], our method explores the synergic effect of cross-scale spatial and frequency contexts

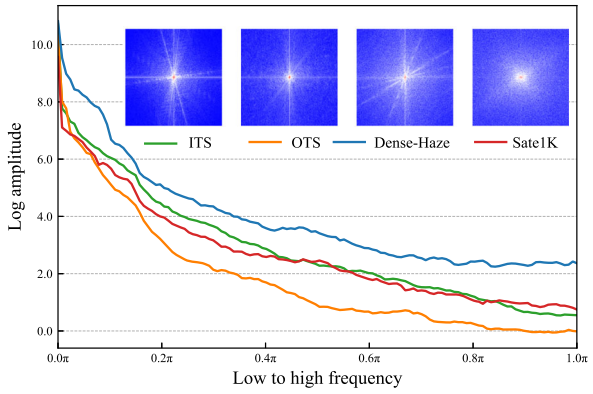

Figure 2 For different image dehazing datasets, such as ITS [20], OTS [20], Dense-Haze [21], and Sate1K [22], we show their log amplitude of Fourier transformed residual images that subtract the hazy images from the clear images. As we can see, there are larger discrepancies in low-frequency components between the paired hazy and clear images

However, these methods still cannot eliminate the local nature of CNNs. Recently, Transformer-based dehazing methods have demonstrated that capturing contextual feature dependencies is vital for haze removal. These include directly employing variants of Swin Transformer [9] into the U-Net framework [10], utilizing Transformer features to modulate convolutional features [11], or applying the Taylor expansion to approximate the softmax-attention [12]. However, these methods are still inefficient due to the quadratic complexity of self-attention. Recent studies have demonstrated that efficient modulation mechanisms [13, 14] can yield satisfactory outcomes with pure convolutional networks and these networks have been shown to be more computationally efficient than Transformers. These methods consider using large kernel convolutional blocks for context modeling and then adopt element-wise multiplication to modulate projected query features, as shown in Fig. 1(a). Despite promising performance improvements in other vision tasks, they overlook the importance of cross-scale context modeling for image dehazing and cannot adaptively extract hierarchical contexts, leading to suboptimal performance.

In addition, the Fourier domain inherently contains global context properties, and rational utilization of Fourier prior information has proven beneficial for dehazing. However, existing methods [15–17] have not sufficiently investigated the varied spectral energy distributions characteristic of hazy images, and consequently lack an effective integration of frequency and spatial modulation mechanisms (see Fig. 1(b)). As shown in Fig. 2, we observe that the difference between the corresponding ground truth and hazy images is primarily in the low-frequency components, indicating that haze significantly affects the low-frequency information. Besides, based on the principle of spectral power distribution in natural images, the power of natural images is statistically concentrated in the low-frequency region [18, 19], suggesting that enhancing the learning of low-frequency features can more effectively eliminate haze degradation and enhance feature representation. Considering the significant variability in the frequency domain among different hazy images, it is necessary to dynamically regulate the low- and high-frequency components to adapt to different samples. In a word, how to adaptively enhance the learning of low-frequency features in the Fourier domain and embed it into the paradigm of the modulation mechanism is a crucial issue.

Taking into account the aforementioned analysis, we introduce a novel modulation mechanism from the perspective of a synergic spatial-frequency domain context. Specifically, in the spatial domain, two consecutive mixture of spatial experts (MSE) modules operating on features at different scales are exploited to learn cross-scale contextual information. For each scale, the MSE employs a multitude of gating experts and a routing network to adaptively select hierarchical features of varying granularity for dynamic learning. In the frequency domain, the mixture of frequency experts (MFE) performs a dynamic selection of frequency context based on multiple frequency experts and a routing network. Each frequency expert consists of two parts: spectral filtering and spectral learning. The former performs mask filtering to enhance the weights of low-frequency components, promoting the removal of haze information; the latter involves Fourier amplitude and separate learning phases, facilitating the global frequency contextual learning. With the assistance of the MoE [23] structure, we adaptively acquire blended dual-domain contexts and then further exploit them to explicitly modulate projected local features, maximizing their collaborative contribution. We call the whole procedure synergic expert modulation (SEM). Finally, based on the proposed module, we construct a basic plug-in block specifically for image dehazing, a synergic expert modulation block, and incorporate it into the U-Net architecture to devise an efficient image dehazing network. We summarize the main contributions as follows:

1) We have devised a novel modulation mechanism called "synergic expert modulation", which integrates two mixture of spatial experts and a mixture of frequency experts to facilitate contextual aggregation and further boost the feature modulation. Leveraging this module, we design an efficient dehazing network.
2) With the assistance of the MoE structure, the MSE is proposed to adaptively aggregate hierarchical contexts, and the MFE is proposed to dynamically regulate low-frequency components and further learn the global frequency context.
3) The proposed method has been evaluated on multiple public benchmarks, demonstrating its superior performance and strong generalization capabilities. Furthermore, it strikes an excellent balance between model complexity and performance.

2 Related work

2.1 End-to-end single image dehazing

Recent research has mainly focused on designing end-to-end models that can directly recover clean images from hazy ones. Many of these models were inspired by or adapted from designs and components in other fields. Ren et al. [24] developed a gated fusion mechanism to combine three inputs, including white balance, contrast enhancement, and gamma correction. Liu et al. [25] implemented attention-based multi-scale estimation on a grid network, enabling adequate information exchange. Dong et al. [5] adopted a simple but effective boosting strategy in the decoder and a back-projection scheme in feature fusion to construct the network. Qin et al. [6] incorporated two local attention mechanisms, channel and pixel attentions, to construct a very deep dehazing network, which can flexibly deal with different types of information. Wu et al. [8] and Zheng et al. [26] utilized the contrastive regularization to reduce the difference between dehazing results and clear label images. More recently, due to the superior long-range context modeling capabilities of Transformers, Refs. [10–12] incorporated variants of self-attention mechanisms into an encoder-decoder architecture to significantly improve model performance. However, these methods often require more computational resources, even if the model capacity is smaller. Liu et al. [27] proposed to utilize a diffusion model and frequency learning for unpaired image dehazing, inspiring other unpaired restoration tasks. Lan et al. [28] adopted stable diffusion as the foundation of the CycleGAN framework, boosting the generalization ability of image dehazing in the real-world scene. However, these methods still cannot handle images in dense haze scenes and are very inefficient. To further improve the efficiency of these methods, we attempt to utilize a pure CNN architecture to fulfill spatial context modulation. Deviating from methods that only consider feature recovery from the spatial domain, Yu et al. [15] and Shen et al. [16, 29] further exploited the integration of features from both the spatial and Fourier domains. However, they fail to explore the unique frequency characteristics of hazy images. Consequently, there is still room for better performance and efficiency trade-offs. Our method considers frequency domain characteristics of haze distribution and customizes a modulation scheme to aggregate frequency context, further achieving synergic feature modulation.

2.2 Mixture-of-experts

The mixture-of-experts (MoE) mechanism [23] operates on a divide-and-conquer principle, distributing learning tasks across multiple specialized neural networks coordinated by a gating mechanism. This architecture produces final predictions through an adaptive combination of outputs from different expert networks. Such dynamic learning capacity further enhances the model's generalization performance across various scenarios. In recent years, the MoE paradigm has been successfully extended to numerous computer vision applications. He et al. [30] incorporated MoE into pan-sharpening to facilitate separate learning of high-frequency and low-frequency components. Cao et al. [31] employed this framework to dynamically fuse local and global representations for infrared and visible image fusion. Yang et al. [32] leveraged vision-language model priors to identify suitable experts for image restoration under diverse weather conditions. These approaches generally only involve the top K experts for certain tasks, following the sparse activation strategy introduced in Ref. [33]. Our work pioneers the application of MoE to image dehazing by comprehensively utilizing each expert's specialized knowledge in both the spatial and frequency domains, which enables context-aware feature aggregation for haze removal.

3 Methods

3.1 Network architecture

Committed to the purpose of efficiency, we adopt a U-Net [34] architecture to construct the overall network, shown in Fig. 3(a). Specifically, in the encoder, we first use a 3×3 convolution to extract the initial hazy features and then stack several synergic expert modulation blocks (SEMBs) to formulate each layer. During the process, the spatial dimension of the feature maps is reduced by half while the channel dimension is increased by twofold. The process of the decoder is opposite to that of the encoder; that is, the spatial dimension gradually increases while the channel dimension gradually decreases. Therefore, the dimensional sizes from the first layer to the fourth layer of the encoder and decoder are $C \times H \times W$, $2C \times \frac{H}{2} \times \frac{W}{2}$, $4C \times \frac{H}{4} \times \frac{W}{4}$, and $8C \times \frac{H}{8} \times \frac{W}{8}$, where H, W, and C represent the height, width, and number of channels of the features, respectively. As depicted in Fig. 3(b), the SEMB

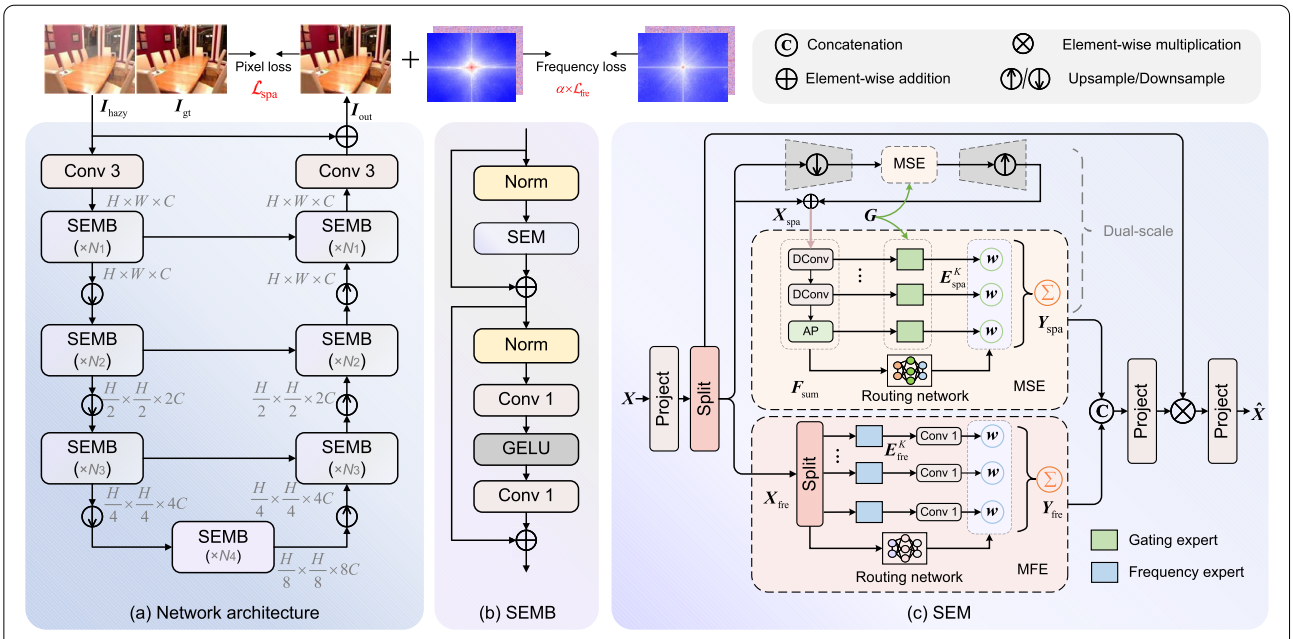

Figure 3 The architecture of the network and its components. (a) The overall pipeline of the proposed synergic expert modulation network (SEMN) approach. $\boldsymbol{I}_{\text{hazy}}$ and $\boldsymbol{I}_{\text{out}}$ represent the corresponding hazy and clear images, respectively. *H*, *W*, and *C* represent the height, width, and number of channels of the features, respectively. (b) The synergic expert modulation block (SEMB). (c) Synergic expert modulation (SEM) first aggregates spatial context using two mixture of spatial experts (MSE) and frequency context using a mixture of frequency experts (MFE) and then incorporates them to modulate projected local features. $\mathcal{L}_{\text{spa}}$ and $\mathcal{L}_{\text{fre}}$ denote spatial loss and frequency loss, respectively. In SEM, $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ denote input and output features, respectively. $\boldsymbol{X}_{\text{spa}}$ and $\boldsymbol{X}_{\text{fre}}$ are input of two branches. $\boldsymbol{F}_{\text{sum}}$ is hierarchical features, $\boldsymbol{G}$ is the gating weights, $\boldsymbol{w}$ is the weight score, $\boldsymbol{E}^k_{\text{spa}}$ is the output of the *k*-th gating expert, and $\boldsymbol{E}^k_{\text{fre}}$ is the output of frequency expert. $\boldsymbol{Y}_{\text{spa}}$ and $\boldsymbol{Y}_{\text{fre}}$ are aggregated spatial and frequency context, respectively

adopts a Transformer-like architecture using synergic expert modulation to model relevant contextual information efficiently and two linear 1×1 convolutions for refined contextual features. To switch the feature scale, we use 4×4 convolution with stride 2 for downsampling and 2×2 deconvolution with stride 2 for upsampling. Similar to most dehazing models, we adopt multiple skip connections in the encoder and the decoder to facilitate the fusion of shallow and deep features. Finally, a convolutional mapping layer is applied to output haze-free images.

3.2 Synergic expert modulation

The focal modulation [14] utilizes the depth-wise convolution (DConv) and a gating mechanism to efficiently aggregate and modulate spatial features, which can be instantiated as

$$\hat{x} = p(ctx(f(x)) \otimes f(x)), \tag{1}$$

where $f(\cdot)$ and $p(\cdot)$ are linear projection layers, $ctx(\cdot)$ is the context extraction function, whose output is a modulator. However, this design fails to address cross-scale feature modulators, which are helpful for the removal of haze degradation. Moreover, haze degradation primarily affects the low-frequency information (see Fig. 2), yet existing methods do not specifically address this issue.

To this end, we propose a synergic expert modulation (SEM) solution, shown in Fig. 3(c), which can capture global contexts in a lightweight manner and utilize them to modulate input query features. The SEM first aggregates spatial and frequency context features based on the proposed mixture of spatial/frequency experts (MSE/MFE), producing dual-domain context modulators, and then interacts with visual query tokens using element-wise multiplication, which can be expressed in Eq. (2):

$$\hat{x} = p(Fuse\big(ctx_{\text{spa}}(f(x)), ctx_{\text{fre}}(f(x))\big) \otimes f(x)), \tag{2}$$

where $ctx_{\text{spa}}(f(x))$ and $ctx_{\text{fre}}(f(x))$ denote spatial context and frequency context aggregation function, respectively, and $Fuse(\cdot)$ is used to integrate extracted context. In particular, we adopt a coarse-to-fine manner to progressively perform spatial context aggregation of features at two scales, that is, utilizing the MSE module on the different resolution features, which can enhance multi-scale representation learning and remove haze at different scales.

3.2.1 Mixture of spatial experts

The MSE consists of two parts: hierarchical feature extraction and feature aggregation. Given the input feature $\boldsymbol{X}_{\text{spa}}$

and gating weights $\boldsymbol{G}$ obtained by performing a channel split operation on projected features, we first stack multiple DConvs with various kernel sizes and a pooling layer to acquire local-to-global features. Each layer output can be denoted as $\boldsymbol{F}_k (k = 1, 2, \ldots, K)$. Then, based on the resulting hierarchical features, we explore utilizing multiple gating experts and a routing network to choose the most suitable context mixture dynamically.

Each gating expert takes the extracted $\boldsymbol{F}_k \in \mathbb{R}^{H\times W\times C}$ and a gating weight $\boldsymbol{G}_k \in \mathbb{R}^{H\times W\times 1}$ as inputs, and then adopts element-wise multiplication to determine how many features need to be aggregated. Thus, the formalization of a gating expert is as follows:

$$\boldsymbol{E}^k_{\text{spa}} = \boldsymbol{F}_k \otimes \boldsymbol{G}_k, \tag{3}$$

where $\boldsymbol{E}^k_{\text{spa}}$ is the output of the k-th gating expert, which will be dynamically integrated based on the routing weights.

To attain routing weights, we first adopt an element-wise addition to integrate all hierarchical features, denoted as $\boldsymbol{F}_{\text{sum}}$, and then input it into the routing network $\mathcal{R}_{\text{spa}}$ to generate probability scores w_{spa} for each spatial expert $\boldsymbol{E}^k_{\text{spa}}$. The routing network is composed of a global average pooling (GAP) layer, a linear layer, and a softmax activation. Therefore, the weight generation process can be defined as follows:

$$\boldsymbol{w}_{\text{spa}} = \text{softmax}(\text{GAP}(\boldsymbol{F}_{\text{sum}}) \cdot \boldsymbol{W}^l_{\text{spa}}), \tag{4}$$

where $\boldsymbol{W}^l_{\text{spa}} \in \mathbb{R}^{C\times K}$ is a learned matrix that maps the pooled features to K expert weights.

Finally, based on the mixed gating expert outputs and obtained weights, the output of the MSE module can be denoted as

$$\boldsymbol{Y}_{\text{spa}} = \sum_{k=1}^{K} \boldsymbol{E}^k_{\text{spa}} \cdot \boldsymbol{w}^k_{\text{spa}}. \tag{5}$$

The output $\boldsymbol{Y}_{\text{spa}}$ is the linearly weighted combination of each gating expert's output with the corresponding routing weight. Such a dynamic learning mechanism can adaptively gather powerful spatial contextual feature representations.

3.2.2 *Mixture of frequency experts*

Given the input feature $\boldsymbol{X}_{\text{fre}}$, the MFE first adopts a channel split operator to divide features uniformly and then deploys multiple frequency expert modules to extract global contexts. Finally, all outputs from experts are aggregated using a specific fusion mechanism.

Frequency expert As shown in Fig. 4, the frequency expert comprises two steps: spectral filtering and spectral learning.

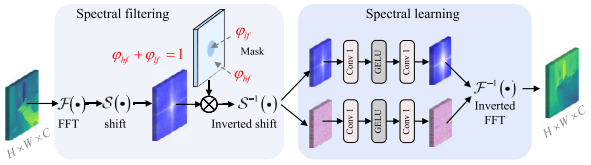

Figure 4 The details of the frequency expert module. FFT denotes fast Fourier transform operation. $\mathcal{F}(\cdot)$ denotes the 2D FFT function, and $\mathcal{S}(\cdot)$ is used for shifting the Fourier information to ensure that low-frequency components are located at the center. $\mathcal{F}^{-1}(\cdot)$ and $\mathcal{S}^{-1}(\cdot)$ denote their inverse operations, respectively

(1) Spectral filtering Research shows that the expected power of the images is statistically concentrated in the low-frequency region. Thus, we consider that low-frequency components should be given higher weights than high-frequency ones. Following Ref. [35], given visual features $\boldsymbol{X}^k_{\text{fre}} \in \mathbb{R}^{H\times W\times C}$, the spectral filtering process can be defined as follows:

$$\tilde{\boldsymbol{X}}^k_{\text{fre}} = \mathcal{F}^{-1}\left(\mathcal{S}^{-1}\left(\boldsymbol{M}_f \otimes \mathcal{S}\left(\mathcal{F}(\boldsymbol{X}^k_{\text{fre}})\right)\right)\right), \tag{6}$$

where $\mathcal{F}(\cdot)$ denotes the 2D fast Fourier transform (FFT) [36] function, and $\mathcal{S}(\cdot)$ is used for shifting the Fourier information to ensure that low-frequency components are located at the center. Correspondingly, $\mathcal{F}^{-1}(\cdot)$ and $\mathcal{S}^{-1}(\cdot)$ denote their inverse operations, respectively. The $\boldsymbol{M}_f$ is a mask map that distributes different weight values on the low- and high-frequency regions. Specifically, φ_{lf} is assigned to the selected low-frequency region $\boldsymbol{A}_{\text{lf}}$, and φ_{hf} is assigned to the rest of the high-frequency part. Hence, the frequency spectral features can be manipulated by adjusting the φ_{lf} values or φ_{hf}:

$$\boldsymbol{M}_f = \begin{cases} \varphi_{\text{lf}} & (u, v) \in \boldsymbol{A}_{\text{lf}} \\ \varphi_{\text{hf}} & \text{otherwise} \end{cases}, \tag{7}$$

where (u, v) is a pair of positions for the low-frequency region, and $\varphi_{\text{lf}} + \varphi_{\text{hf}} = 1$. Therefore, it is vital to determine the boundary of the high- and low-frequency features, that is, to define the range of $\boldsymbol{A}_{\text{lf}}$. Some of previous works [19, 37] adopt a rectangular shape to define it, yet this manner may cause distortion or artifacts in the resulting images. Therefore, we define it as a circular shape in Eq. (8):

$$\boldsymbol{A}_{\text{lf}}(u, v) = \left\{(u, v) \mid (u - u_0)^2 + (v - v_0)^2 < r^2\right\}, \tag{8}$$

where (u_0, v_0) is the origin of (u, v) pairs and r is a radius. In other words, inside the radius and outside the radius, the value of mask $\boldsymbol{M}_f$ is set to φ_{lf} and φ_{hf}, respectively.

(2) Spectral learning Using the mask to modulate frequency feature distribution manually is generally simplistic, yet it neglects the Fourier prior learning. Motivated

by recent work [15], we further conduct Fourier spectral learning after spectral filtering. Specifically, based on the modulated Fourier feature $\boldsymbol{X}_{\mathcal{F}}^{k} = \mathcal{S}^{-1}(\boldsymbol{M}_f \otimes \mathcal{S}(\mathcal{F}(\boldsymbol{X}_{\text{fre}}^{k})))$, we first obtain the corresponding amplitude features $\boldsymbol{X}_{\mathcal{A}}^{k}$ and phase features $\boldsymbol{X}_{\mathcal{P}}^{k}$. Subsequently, we perform two groups of individual convolutional operations, consisting of two 1×1 convolutions and a GELU activation function, on the amplitude and phase features, respectively, to enhance global feature representations. Finally, inverse FFT is used to convert them into the spatial domain. We name the output of each frequency expert as $\boldsymbol{E}_{\text{fre}}^{k}$.

Feature aggregation To accentuate low-frequency information and make the network better adapted to each input sample, it is necessary to specify a reasonable value of mask $\boldsymbol{M}_f$. Here, we adopt mixed frequency experts with various mask values to filter spectra features, which can fully leverage each expert's knowledge to facilitate adaptive frequency modulation. Then, we use a frequency routing network $\mathcal{R}_{\text{fre}}$, similar to the above spatial routing network, to generate probability scores w_{fre} for each expert. The process is calculated using Eq. (9):

$$\boldsymbol{w}_{\text{fre}} = \text{softmax}(\text{GAP}(\boldsymbol{X}_{\text{fre}}) \cdot \boldsymbol{W}_{\text{fre}}^{l}), \tag{9}$$

where $\boldsymbol{W}_{\text{fre}}^{l}$ is the learned weight of linear layer, and $\boldsymbol{w}_{\text{fre}}$ is the obtained weight score. Finally, the aggregation scheme is calculated using Eq. (10):

$$\boldsymbol{Y}_{\text{fre}} = \sum_{k=1}^{K} C_k(\boldsymbol{E}_{\text{fre}}^{k}) \cdot \boldsymbol{w}_{\text{fre}}, \tag{10}$$

where $C_k(\cdot)$ is 1×1 convolution that is used to match the feature dimension of input features, and $\boldsymbol{Y}_{\text{fre}}$ is the aggregated frequency context. In Sect. 4, the detailed routing score of each frequency expert is illustrated in Fig. 10.

3.3 Loss function

Our method adopts a spatial frequency mixed architecture to build the network. Therefore, the loss function consists of two parts: spatial and frequency loss functions. Let $\boldsymbol{I}_{\text{gt}}$ and $\boldsymbol{I}_{\text{out}}$ denote the clear image and dehazed image, respectively, then the loss function can be denoted as

$$\begin{aligned}
\mathcal{L}_{\text{spa}} &= \left\| \boldsymbol{I}_{\text{out}} - \boldsymbol{I}_{\text{gt}} \right\|_1, \\
\mathcal{L}_{\text{fre}} &= \left\| \mathcal{A}\left(\boldsymbol{I}_{\text{out}}\right) - \mathcal{A}\left(\boldsymbol{I}_{\text{gt}}\right) \right\|_1 + \left\| \mathcal{P}\left(\boldsymbol{I}_{\text{out}}\right) - \mathcal{P}\left(\boldsymbol{I}_{\text{gt}}\right) \right\|_1, \\
\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{spa}} + \alpha \mathcal{L}_{\text{fre}},
\end{aligned} \tag{11}$$

where the $\mathcal{L}_{\text{spa}}$ and $\mathcal{L}_{\text{fre}}$ represent spatial and frequency loss, respectively, and $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ denote Fourier amplitude and phase components, respectively; α denotes the trade-off factor, and we empirically set it as 0.05.

4 Experiments

4.1 Setting

Datasets For the image dehazing task, as in the previous work [11], we choose the RESIDE [20] dataset, which includes the Indoor Training Set (ITS) and Outdoor Training Set (OTS), to train our models and evaluate dehazing performance on the SOTS [20] dataset. In particular, SOTS-Indoor and SOTS-Outdoor are used to evaluate the models trained on ITS and OTS, respectively. In addition, the Dense-Haze [21], NH-HAZE [38], and O-HAZE [39] datasets are utilized to evaluate the effectiveness in real-world scenarios, and the Sate1K dataset [22] is used to evaluate the effect of remote sensing dehazing.

Implementation details The number of SEMBs in each layer of the proposed models is set to 2. The channel numbers from the first layer to the fourth layer are 32, 64, 128, and 256. For the MSE, we use three DConvs with various kernel sizes (3, 5, and 7) and a pooling layer to extract features. For the MFE, we use three frequency expert modules with various φ_{lf} values (0.6, 0.7, 0.8, and 0.9). Additionally, we provide a more lightweight version, dubbed SEMN-L, by setting the initial channel numbers to 20. We implement the proposed models on the PyTorch framework with a single NVIDIA 4090 GPU. The ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ is used. Within each mini-batch for different datasets, we augment training samples by applying horizontal or vertical flips and rotations with 90, 180, and 270 degrees.

For the SOTS-Indoor, our models are trained for 800 epochs with an initial learning rate of 2×10^{-4} and then linearly decrease to half every 240 epochs. For the SOTS-Outdoor, our models are trained for 40 epochs with an initial learning rate of 2×10^{-4} and then linearly decrease to half every 10 epochs. The patch and batch sizes are set to 256×256 and 8, respectively.

For Dense-Haze [21], NH-HAZE [38], and O-HAZE [39] datasets, we train all models for a total of 4000 epochs on 800×560 patch size, the initial learning rate is set to 2×10^{-4} and gradually reduced to 2×10^{-6} with the cosine annealing, and the batch size is set to 2.

For the Sate1K [22] dataset, all models are trained for 100 epochs with an initial learning rate of 2×10^{-4} and linearly decay by a factor of 0.95 every 10 epochs. The patch and batch sizes are set to 256×256 and 8.

4.2 Experimental results

We compare dehazing results with several state-of-the-art (SOTA) approaches, including DCP [40], DehazeNet [41], GDN [25], FFA-Net [6], MSBDN [5], AECR-Net [8], Restormer [42], DeHamer [11], SGID-PFF [43], FSDGN [15], MBTFormer-B [12], OKNet [44] and SGDN [45]. The quantitative results on synthetic and real-world datasets are shown in Table 1. Our SEMN approach achieves the best result on the SOTS-Indoor and second-best result

Table 1 Quantitative comparison of state-of-the-art methods for image dehazing. The symbol "-" indicates that the results are unavailable. PSNR and SSIM are evaluation metrics, where higher values indicate better performance. Params and FLOPs are used to evaluate the model's complexity, where lower values indicate better performance. Params: parameters; FLOPs: floating-point operations. The highest score is highlighted in bold

Method	SOTS-Indoor [20]		SOTS-Outdoor [20]		O-HAZE [39]		Dense-Haze [21]		NH-HAZE [38]		Params	FLOPs
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	(M)	(G)
DCP [40]	16.61	0.855	19.14	0.861	16.78	0.653	10.06	0.385	10.57	0.520	-	0.600
DehazeNet [41]	19.82	0.821	24.75	0.927	17.57	0.770	13.84	0.425	16.62	0.524	0.009	0.581
GDN [25]	32.16	0.984	30.86	0.982	18.92	0.672	14.96	0.533	16.46	0.593	0.960	21.50
FFANet [6]	36.59	0.989	33.57	0.984	22.12	0.770	14.39	0.452	19.87	0.692	4.456	287.5
MSBDN [5]	32.77	0.981	34.81	0.986	24.36	0.741	15.37	0.486	19.23	0.706	31.35	41.54
AECR-Net [8]	37.17	0.990	-	-	-	-	14.88	0.505	19.92	0.672	2.611	43.04
Restormer [42]	38.88	0.991	-	-	23.58	0.768	15.78	0.548	-	-	26.10	141.0
DeHamer [11]	36.63	0.988	35.18	0.986	25.11	0.777	16.62	0.560	20.66	0.684	132.4	59.67
SGID-PFF [43]	38.52	0.991	30.20	0.975	20.96	0.741	12.49	0.517	-	-	18.90	81.13
FSDGN [15]	38.63	0.990	-	-	-	-	16.91	0.581	19.99	**0.731**	2.731	19.59
MBTFormer-B [12]	40.71	0.992	37.42	0.989	25.05	0.788	16.66	0.560	-	-	2.662	38.50
OKNet [44]	40.79	0.993	**37.68**	0.989	25.64	0.784	16.92	0.608	20.48	0.712	4.720	39.67
SGDN [45]	40.41	0.889	37.25	0.985	25.04	0.743	16.56	0.593	20.02	0.691	5.41	29.50
SEMN-L (ours)	39.97	0.993	36.51	0.989	25.10	0.781	17.03	0.584	20.39	0.655	1.386	6.481
SEMN (ours)	**41.46**	**0.995**	37.60	**0.991**	**25.94**	**0.794**	**17.52**	**0.607**	**20.70**	0.700	3.379	15.52

Table 2 Quantitative comparison on the Sate1K [22] dataset for non-uniform satellite image haze removal

Method	Thin haze		Moderate haze		Thick haze	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DeHamer [11]	22.77	0.933	26.37	0.943	22.37	0.899
FSDGN [15]	21.73	0.946	25.77	0.970	22.35	0.937
DCINet [46]	20.19	0.947	**27.43**	0.964	21.45	0.926
TrinityNet [47]	21.30	0.946	26.47	0.963	20.76	0.915
SEMN (ours)	**23.44**	**0.960**	26.64	**0.972**	**23.79**	**0.946**

on the SOTS-Outdoor dataset, achieving 41.46 dB and 37.60 dB PSNR scores, outperforming MBTFormer-B by 0.75 dB and 0.18 dB, respectively. As a lightweight version, SEMN-L performs equally well. It can achieve better results than FSDGN with only 51% of the parameters and 33% of FLOPs. Furthermore, except for the SSIM value on the NH-HAZE dataset, our method outperforms most other dehazing methods for real-world datasets. For the remote sensing dehazing, the quantitative results are shown in Table 2. It is evident from these results that our method continues to perform well on the Sate1K [22] dataset with different haze densities.

We provide visual examples of the SOTS dataset in Figs. 5 and 6, from which we observe that the proposed SEMN approach can effectively eliminate haze and restore satisfactory details and textures. The visualizations from real-world datasets, Dense-Haze and NH-HAZE, are shown in Fig. 7. We find that images produced by our method do not exhibit severe color casts or artifacts. Conversely, other methods produce unexpected results.

4.3 Model complexity analysis

We summarize the model parameters and floating-point operations (FLOPs) in Table 1. Compared to the second-best Transformer-based method, MBTFormer-B [12], our SEMN approach achieves a superior balance between computation and performance. In comparison to the Fourier-based FSDGN [15], our lightweight variant SEMN-L consumes only half the parameters and computational cost while achieving better results. In addition, the inference time comparisons of several representational Fourier and Transformer-based methods are shown in Fig. 8, where the results are an average value obtained after ten repeated experiments. Distinctly, it can be observed that although the inference speed of our method is lower compared to DeHamer [11] and FSDGN, the performance is significantly improved. The MBTFormer-B and DehazeFormer-L [10] can obtain more than 40 dB PSNR scores, but the high computational cost of self-attention hinders their efficiency. These results suggest that our methods achieve high performance and are more practical and efficient.

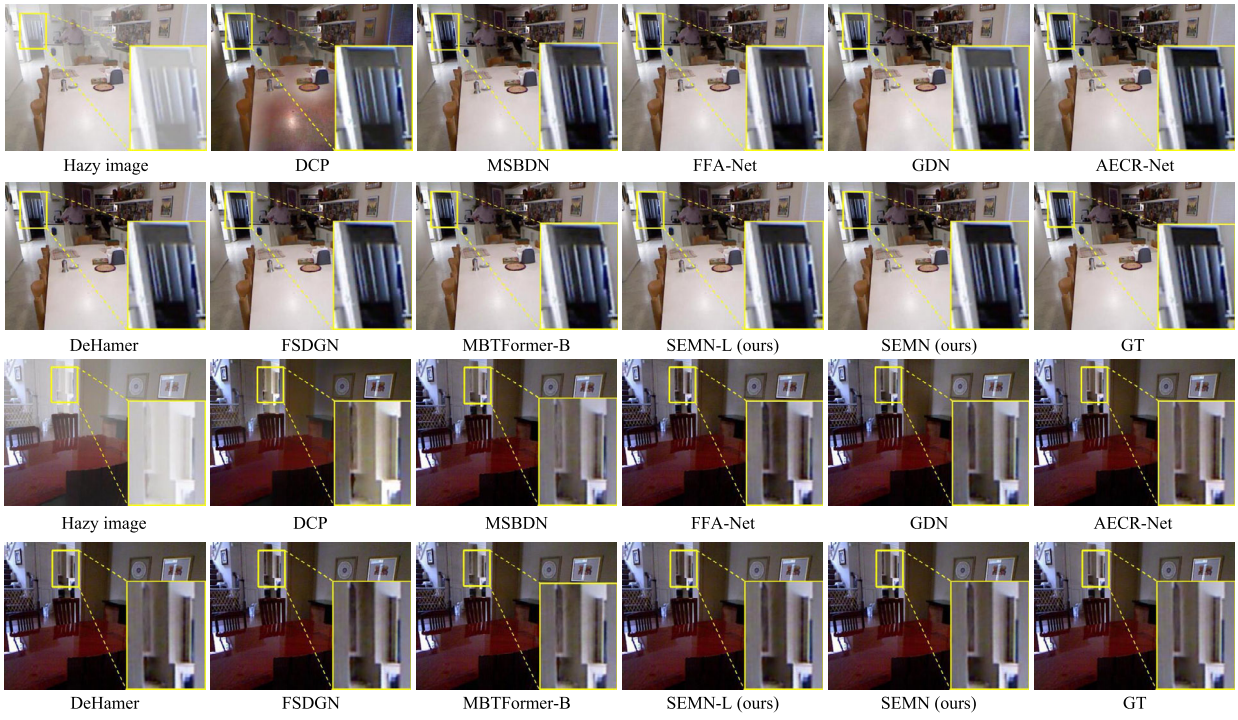

Figure 5 Visual comparisons on the SOTS-Indoor [20] dataset

Figure 6 Visual comparisons on the SOTS-Outdoor [20] dataset

4.4 Ablation studies

We conduct ablation studies to verify the proposed components. All experiments are performed on the SEMN, trained on the ITS dataset with 400 epochs, and evaluated on the SOTS-Indoor dataset.

Synergic expert modulation We evaluate the effectiveness of the proposed SEM in Table 3. The baseline model is implemented by replacing the SEMB with a basic residual block. The results show an obvious performance improvement after substituting it with our proposed block. Specifically, since the SEM contains two components, MSE and MFE, we conduct separate experiments to verify them. The second row means we solely add the MSE in the SME module, and the PSNR shows 2.02 dB performance gains compared to the baseline. The MFE mainly operates in the Fourier domain, so we further combine it with MSE to con-

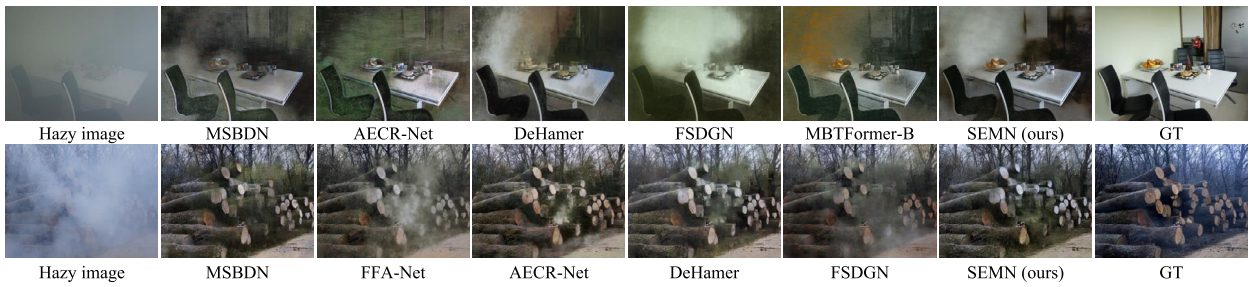

Figure 7 Visual comparisons on Dense-Haze [21] and NH-HAZE [38] datasets

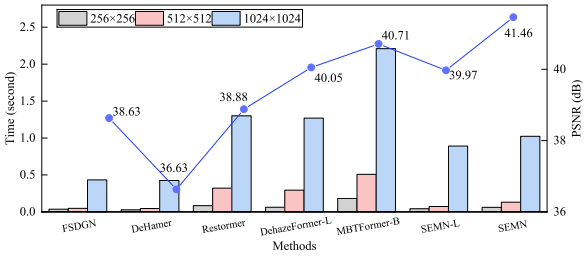

Figure 8 The inference time comparisons of SOTA dehazing methods on 256 × 256, 512 × 512 and 1024 × 1024 images

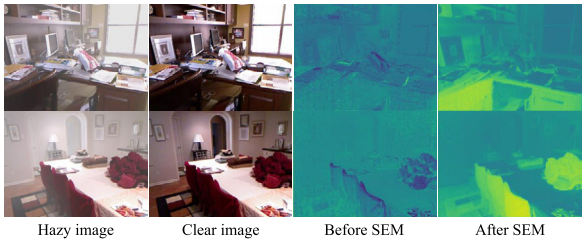

Figure 9 Feature visualization. We show visualizations of feature maps before and after employing SEM

Table 3 Ablation study on the synergic expert modulation (SEM). FLOPs are computed on the image patch size of 256 × 256 × 3

Model	MSE	MFE	PSNR (dB)	Params (M)	FLOPs (G)
Baseline	✗	✗	37.41	4.51	19.94
SEMN (ours)	✓	✗	39.43	2.71	12.68
	✓	✓	40.63	3.38	15.52

Table 4 Ablation study of individual components on the mixture of spatial experts (MSE). FLOPs are computed on the image patch size of 256 × 256 × 3

Model	PSNR (dB)	Params (M)	FLOPs (G)
MSE w/o gating expert	39.94	3.38	15.52
MSE w/o routing network	40.02	3.37	15.50
MSE w/o both	39.66	3.37	15.50
SEMN (ours)	40.63	3.38	15.52

Table 5 Ablation study of individual components on the mixture of frequency experts (MFE) . FLOPs are computed on the image patch size of 256 × 256 × 3

Model	PSNR (dB)	Params (M)	FLOPs (G)
MFE w/o frequency expert	39.74	3.16	14.58
MFE w/o routing network	40.11	3.37	15.50
MFE w/o both	39.57	3.15	14.56
SEMN (ours)	40.63	3.38	15.52

struct the SEM module in the third row. We note that, the model achieves the best result, which demonstrates the feasibility and effectiveness of the synergic context modeling mechanism. Additionally, we visualize the features before and after integrating the SEM into the SEMB as depicted in Fig. 9. When applying the SEM, the features contain sharper textures and details, presenting superior context modulation capabilities.

Mixture-of-experts mechanism The core designs of the mixture of spatial experts (MSE) and the mixture of frequency experts (MFE) are the corresponding experts and the routing network, so we carry out breakdown ablation to investigate their efficacy in Tables 4 and 5. It is particularly emphasized that the output is the sum of all experts when the routing network is not used. Firstly, in the MSE, when two components are removed separately, the corresponding models drop 0.69 dB and 0.61 dB PSNR values, respectively, compared with our SEMN. Besides, the performance is severely degraded when both modules are discarded. This demonstrates the adaptive learning capabilities of gating experts and the dynamic feature aggregation capabilities of routing networks. Moreover, the design does not result in excessive model parameters and model complexity, which benefits the building of efficient dehazing networks. Secondly, in the MFE, we can find that all the results after removing modules appear to have significantly decreased, verifying the importance of frequency experts and the routing mechanism.

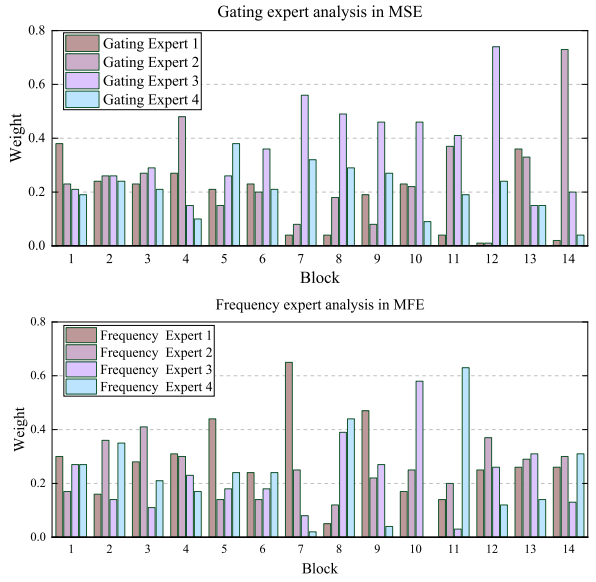

Figure 10 The analysis of the mixture-of-experts (MoE) mechanism. We plot the weight scores assigned to each gating expert and frequency expert of SEMB

Table 6 Ablation study for the dual-scale design in SEM

Model	PSNR (dB)	Params (M)	FLOPs (G)
Single-scale	40.16	3.26	15.21
Dual-scale (ours)	40.63	3.38	15.52

To better grasp the MoE mechanism, we present the routing weights of all 14 SEMBs allocated to various gating and frequency experts in Fig. 10. For gating experts, due to the dual-scale design, we only present the results of the second MSE in each SEMB. In our design, gating experts aggregate hierarchical features obtained using DConvs with different kernel sizes (3, 5, and 7) or global pooling operations. Therefore, the gating expert can see that earlier blocks show a diverse range of expert choices, meaning that extracting low-level features requires various hierarchical features. However, deep layers tend to reconstruct structural features, and the convolution with large kernels can better achieve the target. With respect to the frequency experts, there is no regularity in the routing weights. Therefore, it can be seen that if we set a uniform value for $\boldsymbol{M}_f$ in the frequency domain expert and without the routing network, the model hardly dynamically adapts to different samples. These phenomena also demonstrate the rationality of our design.

Dual-scale design of SEM We employ two MSE modules to build SEM. Therefore, we conduct experiments to verify the effectiveness of the dual-scale design (See Table 6). When equipped with SEM, the model's performance increases by 0.47 dB with only a slight increase in model parameters and computational cost. Figure 11 visualizes hierarchical features and output features at each scale, and we observe low-scale features focusing more on abstract context and high-scale features extracting sharper textures and edges. This further demonstrates that the dual-scale design can provide more powerful representations.

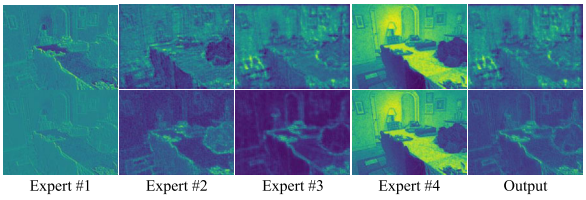

Figure 11 Visualization of the features learned by each gating expert in MSE. The first and second rows are the features operating on low-scale and high-scale space, respectively

Table 7 Ablation study of frequency expert (FE). FLOPs are computed on the image patch size of 256 × 256 × 3

Model	PSNR (dB)	Params (M)	FLOPs (G)
FE w/o spectral filtering	39.82	3.38	15.52
FE w/o spectral learning	40.15	3.16	14.58
SEMN (ours)	40.63	3.38	15.52

Frequency expert To investigate the importance of components within the frequency expert, we start by removing the spectral filtering operation and observe a performance degradation of 0.81 dB. Next, we eliminate the spectral learning process, resulting in a 0.48 dB drop in performance (See Table 7). These findings indicate that adjusting and learning the Fourier spectrum can significantly enhance dehazing performance. Furthermore, we examine the impact of specific parameters in spectral filtering on performance, as shown in Table 8. Firstly, we set φ_{lf} to less than 0.5 to emphasize high-frequency components. There is a cliff-like drop in performance, down by 1.52 dB, which is consistent with our motivation that assigning high weights to low-frequency components is beneficial for haze removal. Secondly, we probe into the influence of the radius value of the low-frequency region on the performance. Setting $r = 2$ is marginally better than $r = 4$; therefore, we default to $r = 2$ for constructing all networks.

In Table 9, we investigate an additional set of experiments, selecting 1 to 4 frequency experts higher than 0.5 for each experiment. We found that the performance was optimal when the value was set to 4.

5 Conclusion

This paper proposes an efficient synergic expert modulation network (SEMN) for image dehazing. The core design is the proposed synergic expert modulation (SEM) mechanism, which fully leverages dual-domain contextual information to modulate initial local features. In detail, the SEM

Table 8 Parameter settings in spectral filtering

Model	PSNR (dB)	Params (M)	FLOPs (G)
$\varphi_{\text{lf}} = [0.1, 0.2, 0.3, 0.4]$	39.11	3.38	15.52
r (radius) = 4	40.32	3.38	15.52
r (radius) = 2 (ours)	40.63	3.38	15.52

Table 9 Ablation study on the number of frequency experts

Model	PSNR (dB)
N = 1: $\varphi_{\text{lf}} = [0.6]$	40.09
N = 2: $\varphi_{\text{lf}} = [0.6, 0.7]$	40.37
N = 3: $\varphi_{\text{lf}} = [0.6, 0.7, 0.8]$	40.41
N = 4: $\varphi_{\text{lf}} = [0.6, 0.7, 0.8, 0.9]$	40.63

comprises two core modules: the mixture of spatial experts (MSE) and the mixture of frequency experts (MFE). Both modules employ a mixture of experts (MoE) mechanism to integrate specialized knowledge from various experts of dual domains dynamically. The MSE adaptively learns hierarchical features of varying granularity, guided by multiple gating experts and a routing network. At the same time, the MFE focuses on mining frequency context guided by multiple frequency experts and a routing network. Based on this design, the basic building block named SEMB is devised. Equipping it into the U-Net architecture, the proposed SEMN approach can achieve superior dehazing performance with acceptable model complexity.

Abbreviations

CNN, convolutional neural network; GT, ground truth; MoE, mixture-of-experts; MFE, mixture of frequency experts; MSE, mixture of spatial experts; SEM, synergic expert modulation; SEMB, synergic expert modulation block; SEMN, synergic expert modulation network; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index measure.

Author contributions

All authors contributed to the study conception and design. HD and YZ provided technical and theoretical support for the research. XC handled the data preparation and survey analysis. Method design, data collection, and analysis were performed by HS. HS wrote the first draft of the manuscript, which was revised by XJ. All authors read and approved the final manuscript.

Funding information

This work was supported by the Scientific Research Foundation for High-level Talent of Anhui University of Science and Technology (No. 2025yjrc0015), and the National Natural Science Foundation of China (Nos. 62502006 and 62472104).

Data availability

The datasets generated during and/or analyzed during the current study are available at https://github.com/it-hao/. The detailed information is as follows: RESIDE: https://sites.google.com/view/reside-dehaze-datasets/reside-standard?authuser=3D0; Dense-Haze: https://data.vision.ee.ethz.ch/cvl/ntire19//dense-haze/; NH-HAZE: https://data.vision.ee.ethz.ch/cvl/ntire20/nh-haze/; O-HAZE: https://data.vision.ee.ethz.ch/cvl/ntire18/o-haze/; Sate1K: https://drive.google.com/drive/folders/1eeBA2V_I9-evSJ0XWhRAww6ftweq8hU_.

Declarations

Competing interests

Henghui Ding and Yulun Zhang are Associate Editors at Visual Intelligence and were not involved in the editorial review of this article or the decision to publish it. The authors declare that they have no other competing interests.

Author details

[1]State Key Laboratory of Digital Intelligent Technology for Unmanned Coal Mining, Anhui University of Science and Technology, Huainan, 232001, China. [2]School of Public Security and Emergency Management, Anhui University of Science and Technology, Hefei, 231131, China. [3]School of Cyber Science and Engineering, Southeast University, Nanjing, 210096, China. [4]Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, 200433, China. [5]AI Institute, Shanghai Jiao Tong University, Shanghai, 200240, China. [6]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, Singapore.

Received: 30 October 2025 Revised: 16 January 2026
Accepted: 18 January 2026 Published online: 10 February 2026

References

1. Liu, Y., Zhao, G., Gong, B., Li, Y., Raj, R., Goel, N., Kesav, S., Gottimukkala, S., Wang, Z., Ren, W., et al. (2018). Improved techniques for learning to dehaze and beyond: a collective study. arXiv preprint. arXiv:1807.00202.
2. Tan, R. T. (2008). Visibility in bad weather from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1–8). Piscataway: IEEE.
3. Sakaridis, C., Dai, D., Hecker, S., & Van Gool, L. (2018). Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Proceedings of the European conference on computer vision* (pp. 687–704). Cham: Springer.
4. Ren, W., Zhang, J., Xu, X., Ma, L., Cao, X., Meng, G., & Liu, W. (2018). Deep video dehazing with semantic segmentation. *IEEE Transactions on Image Processing*, *28*, 1895–1908.
5. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., & Yang, M.-H. (2020). Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2157–2167). Piscataway: IEEE.
6. Qin, X., Wang, Z., Bai, Y., Xie, X., & Jia, H. (2020). FFA-Net: feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11908–11915). Palo Alto: AAAI Press.
7. Chen, Z., He, Z., & Lu, Z.-M. (2024). DEA-Net: single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, *33*, 1002–1015.
8. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., & Ma, L. (2021). Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10551–10560). Piscataway: IEEE.
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022). Piscataway: IEEE.
10. Song, Y., He, Z., Qian, H., & Du, X. (2023). Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, *32*, 1927–1941.
11. Guo, C.-L., Yan, Q., Anwar, S., Cong, R., Ren, W., & Li, C. (2022). Image dehazing transformer with transmission-aware 3D position embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5812–5820). Piscataway: IEEE.
12. Qiu, Y., Zhang, K., Wang, C., Luo, W., Li, H., & Jin, Z. (2023). MB-TaylorFormer: multi-branch efficient transformer expanded by Taylor formula for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12802–12813). Piscataway: IEEE.
13. Ma, X., Dai, X., Yang, J., Xiao, B., Chen, Y., Fu, Y., & Yuan, L. (2025). Efficient modulation for vision networks. In *Proceedings of the international conference on learning representations* (pp. 1–19). Retrieved April 24, 2025, from https://openreview.net/forum?id=ip5LHJs6QX.
14. Yang, J., Li, C., Dai, X., & Gao, J. (2022). Focal modulation networks. In *Proceedings of the international conference on neural information processing systems* (pp. 4203–4217). Red Hook: Curran Associates.
15. Yu, H., Zheng, N., Zhou, M., Huang, J., Xiao, Z., & Zhao, F. (2022). Frequency and spatial dual guidance for image dehazing. In S. Avidan, G. Brostow, M.

Cissé, G. M. Farinella, & T. Hassner (Eds.), *Proceedings of the European conference on computer vision* (pp. 181–198). Cham: Springer.
16. Shen, H., Zhao, Z.-Q., Zhang, Y., & Zhang, Z. (2023). Mutual information-driven triple interaction network for efficient image dehazing. In *Proceedings of the ACM international conference on multimedia* (pp. 7–16). New York: ACM.
17. Zhou, M., Huang, J., Guo, C.-L., & Li, C. (2025). Fourmer: an efficient global modeling paradigm for image restoration. In *Proceedings of the international conference on machine learning* (pp. 42589–42601). Retrieved April 25, 2025, from https://openreview.net/forum?id=9lvywx8c2t.
18. Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412.
19. Rippel, O., Snoek, J., & Adams, R. P. (2015). Spectral representations for convolutional neural networks. In *Proceedings of the 29th international conference on neural information processing systems* (pp. 2449–2457). Red Hook: Curran Associates.
20. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., & Wang, Z. (2018). Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, *28*, 492–505.
21. Ancuti, C. O., Ancuti, C., Sbert, M., & Timofte, R. (2019). Dense-haze: a benchmark for image dehazing with dense-haze and haze-free images. In *Proceedings of the IEEE international conference on image processing* (pp. 1014–1018). Piscataway: IEEE.
22. Huang, B., Zhi, L., Yang, C., Sun, F., & Song, Y. (2020). Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks. In *Proceedings of the IEEE winter conference on applications of computer vision* (pp. 1806–1813). Piscataway: IEEE.
23. Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review*, *42*, 275–293.
24. Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., & Yang, M.-H. (2018). Gated fusion network for single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3253–3261). Piscataway: IEEE.
25. Liu, X., Ma, Y., Shi, Z., & Chen, J. (2019). GridDehazeNet: attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7314–7323). Piscataway: IEEE.
26. Zheng, Y., Zhan, J., He, S., Dong, J., & Du, Y. (2023). Curricular contrastive regularization for physics-aware single image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5785–5794). Piscataway: IEEE.
27. Liu, C., Qi, L., Pan, J., Qian, X., & Yang, M.-H. (2025). Frequency domain-based diffusion model for unpaired image dehazing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1–8). Piscataway: IEEE.
28. Lan, Y., Cui, Z., Liu, C., Peng, J., Wang, N., Luo, X., & Liu, D. (2025). Exploiting diffusion prior for real-world image dehazing with unpaired training. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 4455–4463). Palo Alto: AAAI Press.
29. Shen, H., Ding, H., Zhang, Y., Zhao, Z.-Q., & Jiang, X. (2025). Spatial frequency modulation network for efficient image dehazing. *IEEE Transactions on Image Processing*, *34*, 3982–3996.
30. He, X., Yan, K., Li, R., Xie, C., Zhang, J., & Zhou, M. (2024). Frequency-adaptive pan-sharpening with mixture of experts. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2121–2129). Palo Alto: AAAI Press.
31. Cao, B., Sun, Y., Zhu, P., & Hu, Q. (2023). Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 23555–23564). Piscataway: IEEE.
32. Yang, H., Pan, L., Yang, Y., & Liang, W. (2024). Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 24902–24912). Piscataway: IEEE.
33. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. arXiv preprint. arXiv:1701.06538.
34. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In *Proceedings of the 18th international conference on medical image computing and computer-assisted intervention* (pp. 234–241). Cham: Springer.
35. Yun, G., Yoo, J., Kim, K., Lee, J., & Kim, D. H. (2023). SPANet: frequency-balancing token mixer using spectral pooling aggregation modulation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6113–6124). Piscataway: IEEE.
36. Brigham, E. O., & Morrow, R. (1967). The fast Fourier transform. *IEEE Spectrum*, *4*, 63–70.
37. Zhou, M., Huang, J., Li, C., Yu, H., Yan, K., Zheng, N., & Zhao, F. (2022). Adaptively learning low-high frequency information integration for pan-sharpening. In *Proceedings of the ACM international conference on multimedia* (pp. 3375–3384). New York: ACM.
38. Ancuti, C. O., Ancuti, C., & Timofte, R. (2020). NH-HAZE: an image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshop* (pp. 444–445). Piscataway: IEEE.
39. Ancuti, C. O., Ancuti, C., Timofte, R., & De Vleeschouwer, C. (2018). O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshop* (pp. 754–762). Piscataway: IEEE.
40. He, K., Sun, J., & Tang, X. (2010). Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*, 2341–2353.
41. Cai, B., Xu, X., Jia, K., Qing, C., & Tao, D. (2016). DehazeNet: an end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, *25*, 5187–5198.
42. Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5728–5739). Piscataway: IEEE.
43. Bai, H., Pan, J., Xiang, X., & Tang, J. (2022). Self-guided image dehazing using progressive feature fusion. *IEEE Transactions on Image Processing*, *31*, 1217–1229.
44. Cui, Y., Ren, W., & Knoll, A. (2024). Omni-kernel modulation for universal image restoration. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*, 12496–12509.
45. Fang, W., Fan, J., Zheng, Y., Weng, J., Tai, Y., & Li, J. (2025). Guided real image dehazing using YCbCr color space. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2906–2914). Palo Alto: AAAI Press.
46. Zhang, L., & Wang, S. (2022). Dense haze removal based on dynamic collaborative inference learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–16.
47. Chi, K., Yuan, Y., & Wang, Q. (2023). Trinity-Net: gradient-guided swin transformer-based remote sensing image dehazing and beyond. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–14.

Publisher's note