Beyond Spatial Domain: Cross-domain Promoted Fourier Convolution Helps Single Image Dehazing

Xiaozhe Zhang[1,2], Haidong Ding[1,2], Fengying Xie[1,2*], Linpeng Pan[1,2], Yue Zi[3], Ke Wang[1,2], Haopeng Zhang[1,2]

[1]School of Astronautics, Beihang University
[2]Tianmushan Laboratory, Beihang University
[3]School of Electrical and Information Engineering, Changsha University of Science and Technology
{xiaozhe_zhang, dinghaidong, xfy_73, linpengpan, wang_ke, zhanghaopeng}@buaa.edu.cn, ziyue91@csust.edu.cn

Abstract

Vanilla convolution and window-based self-attention have shown significant success in image dehazing. However, they are constrained by limited receptive fields and ignore frequency gaps between dehazed and clear images. The former hampers the modeling of global dependencies, while the latter impedes the learning of high-frequency features, leading to suboptimal performance. In this paper, we propose the Joint Spatial and Fourier Convolutional Network (JSFC-Net), which leverages Fourier transformation to simultaneously address the two aforementioned problems with low computational overhead. We introduce the Frequency-Spatial Promoted and Physical Learning Block, which extracts high-level features from the spatial domain and frequency domain in parallel. We design a simple yet effective solution that uses spatial features to promote and modulate frequency features in a multi-scale manner, achieving refinement of frequency features and addressing robustness issue caused by global sensitivity. Additionally, we present the Receptive Field Selection Module to facilitate improved fusion of spatial and frequency domain features. Finally, we introduce frequency loss to further narrow frequency gaps. Comprehensive experiments on multiple datasets demonstrate that JSFC-Net is significantly superior to SOTA dehazing methods.

Introduction

Images captured under hazy conditions often suffer from reduced visibility and low contrast due to atmospheric particles (Nayar and Narasimhan 1999). These degraded images have limited usability in subsequent high-level computer vision tasks, such as object detection (Li et al. 2023) and semantic segmentation (Ren et al. 2018b). As a result, image dehazing has become a crucial pre-processing step.

Existing dehazing methods can be classified into two main categories: prior-based methods (Fattal 2008; He, Sun, and Tang 2010; Fattal 2014) and data-driven methods (Ren et al. 2020; Cai et al. 2016; Li et al. 2017; Zhang and Patel 2018; Ren et al. 2018a; Dong et al. 2020; Qin et al. 2020; Song et al. 2023; Wang et al. 2023; Zhang et al. 2024; Kulkarni, Phutke, and Murala 2022). Prior-based methods are based on the Atmospheric Scattering Model (ASM) (McCartney 1976; Narasimhan and Nayar 2002), incorporating various physical assumptions regarding image statistics. However, these methods often perform poorly when the statistical priors do not hold in real-world scenarios.

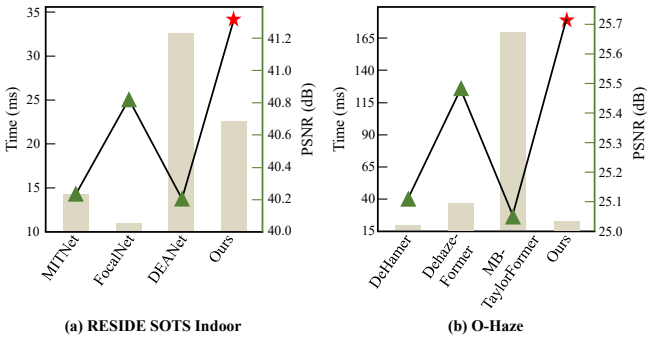


Figure 1: Comparison of performance and efficiency trade-offs with SOTA methods on (a) RESIDE SOTS Indoor and (b) O-Haze datasets. The line graph represents PSNR, and the bar graph represents inference time.

In contrast, data-driven methods generally yield better results. However, several limitations and challenges remain. ***1) Receptive Fields.*** Data-driven methods typically employ vanilla convolution or window-based self-attention to extract features. However, these local operators have limited receptive fields. Previous work has either relied on larger and deeper models or adopted an encoder-decoder architecture to address this issue. However, simply increasing the capacity of deep models does not consistently lead to improved performance (Kong et al. 2023). Moreover, reducing the spatial resolution of features may result in information loss. Most importantly, none of them have direct access to the global receptive field in a single block. ***2) Restoration of High-Frequency Details.*** Dehazing methods aim to recover two components of haze-free images: low-frequency structures and high-frequency details. However, learning both components simultaneously is challenging for models that operate solely in the spatial domain (Pan et al. 2022). While many existing dehazing methods successfully preserve the consistency of low-frequency structures, they struggle to restore high-frequency details effectively.

To simultaneously address challenges mentioned above, we propose the **J**oint **S**patial and **F**ourier **C**onvolutional

*Corresponding author.

Network (JSFC-Net). Drawing inspiration from Fourier transform theory (Katznelson 2004), our key insight is that altering a single value in the frequency domain has a global impact on the Fourier transform-involved input features. Therefore, we introduce convolution operations on the frequency domain features, enabling the acquisition of global receptive fields. Additionally, Fourier convolution allows for more direct learning of high-frequency features, which aids in restoring high-frequency details. Given that the Fourier spectrum can be efficiently computed using the Fast Fourier Transform (FFT), with a computational complexity of $\mathcal{O}((NM)\log(NM))$ for an image of size $N \times M$, JSFC-Net addresses the two aforementioned issues with low computational overhead. Furthermore, our approach no longer relies on larger models or encoder-decoder architectures, avoiding performance bottlenecks and the loss of spatial information.

The core building block of JSFC-Net is the well-designed Frequency-Spatial Promoted and Physical Learning (FSPPL) Blocks, which consist of a Frequency Promoted Learning Branch (FPLB), a Spatial Physical Learning Branch (SPLB), and a Receptive Field Selection Module (RFSM). The SPLB enforces physics-based priors from the ASM in the feature space, capturing interpretable local spatial features. The FPLB extracts global frequency features by performing convolution operations in frequency domain. To further refine frequency features and address robustness issues caused by global sensitivity, we propose a simple yet effective strategy that utilizes spatial features to promote and modulate frequency features in a multi-scale manner, while employing a Mixture-of-Experts (MoE) approach to aggregate the results of multi-scale modulation. Finally, we use the RFSM to predict the most suitable receptive field size for each pixel, ensuring better fusion of the two domains. Additionally, we introduce a frequency loss that enhances the quality of dehazed images by reducing frequency gaps. Comprehensive quantitative and qualitative experimental results demonstrate the superiority of JSFC-Net over SOTA methods. And we are the first to demonstrate that frequency loss can serve as a universal regularization technique to enhance the dehazing performance of various SOTA methods.

Our contributions can be summarized as follows:

- We propose the Joint Spatial and Fourier Convolutional Network (JSFC-Net). It skilfully utilizes Fourier Transform to achieve the global receptive field and addresses the difficulty of high-frequency learning in existing dehazing models with low computational complexity.
- We develop a Frequency-Spatial Promoted and Physical Learning Block that effectively facilitates the complementarity between spatial and frequency domain features through cross-domain promotion and adaptive fusion.
- Extensive experiments demonstrate that JSFC-Net significantly outperform existing SOTA methods in various benchmarks, particularly in real-world dehazing.

Related Work

Single Image Dehazing

Single image dehazing is a highly ill-posed problem, so most early prior-based methods (Fattal 2008; He, Sun, and Tang 2010; Fattal 2014; Zhu, Mai, and Shao 2015) employ various priors or assumptions to estimate crucial parameters in the ASM. For example, He *et al.* (He, Sun, and Tang 2010) introduced the dark channel prior based on statistical laws, exploiting the observation that pixel values in non-sky areas tend to approach zero. However, these methods encounter difficulties when applied to real-world images where the underlying physical assumptions are invalid.

In early data-driven methods, networks were designed to estimate key parameters within the ASM. Cai *et al.*(Cai et al. 2016) proposed a trainable framework that utilizes Bilateral Rectified Linear Units for estimating intermediate transmission. Recently, ASM-independent deep networks (Ren et al. 2018a; Liu et al. 2019; Dong et al. 2020; Zhang et al. 2020; Qin et al. 2020; Wu et al. 2021; Song et al. 2023; Yu et al. 2022) have been introduced to directly estimate clear images or haze residuals. FFA-Net, introduced by Qin *et al.* (Qin et al. 2020), is a representative CNN-based model that assigns varying importance to different features, thereby enhancing the representational capacity. DehazeFormer (Song et al. 2023) is the first dehazing network to use the Swin Transformer (Liu et al. 2021) as its backbone, incorporating several key design modifications such as normalization layers, activation functions, and spatial information aggregation. FSDGN (Yu et al. 2022) is the first work to apply frequency-domain processing to image dehazing. It reconstructs the phase spectrum under the guidance of amplitude spectrum and integrates global frequency information to facilitate local feature learning in the spatial domain.

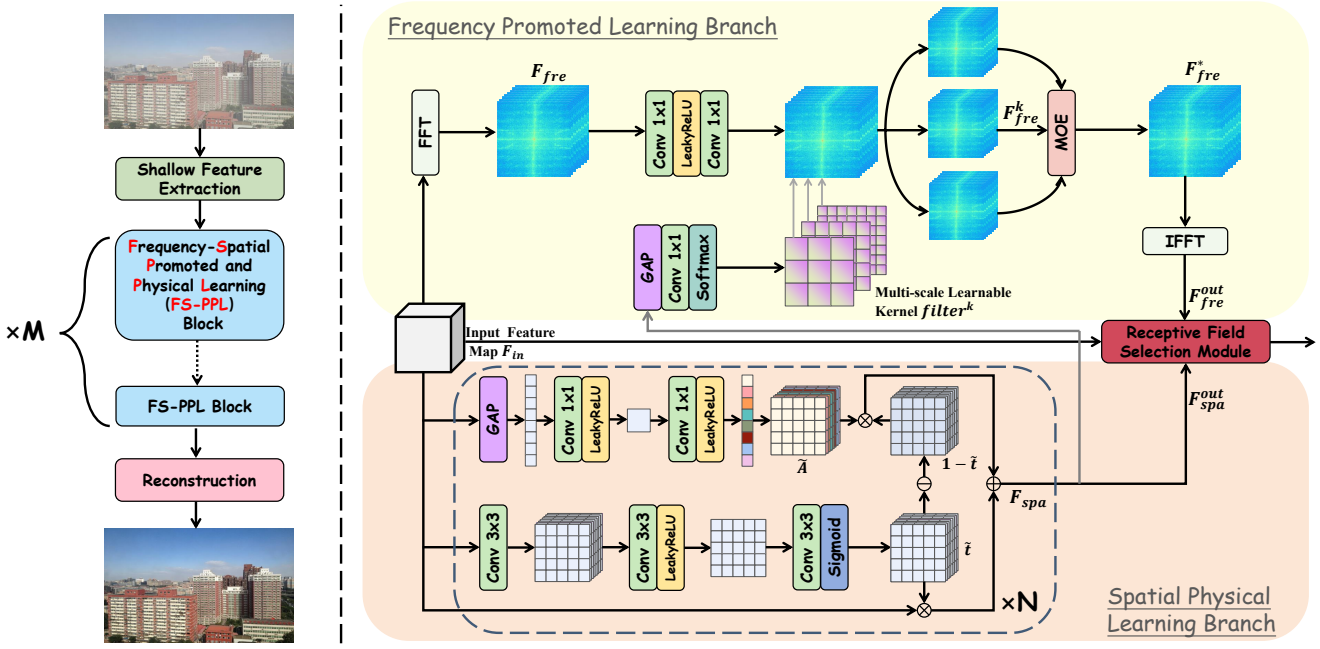
Frequency Learning in Vision Tasks

With the breakthrough of deep learning technology, several researchers (Chi, Jiang, and Mu 2020; Rippel, Snoek, and Adams 2015; Jiang et al. 2021; Zhou et al. 2022; Cui et al. 2023b; Miao, Deng, and Han 2024) have explored incorporating frequency analysis into deep neural networks for vision tasks. For example, Chi *et al.*(Chi, Jiang, and Mu 2020) replaced traditional convolutions with a Fourier Unit to achieve non-local receptive fields in deep models. Rao *et al.*(Rao et al. 2021) proposed a computationally efficient architecture that substitutes the self-attention layer in the Vision Transformer with a 2D discrete Fourier transform in the frequency domain. Additionally, several studies (Rippel, Snoek, and Adams 2015; Xu et al. 2020; Wang et al. 2016) have utilized frequency analysis to compress networks and accelerate CNNs. Recent applications of frequency-domain analysis in vision tasks have achieved remarkable success, including image recognition (Frank et al. 2020), style transfer (Yoo et al. 2019), and super-resolution (Wei et al. 2021).

Proposed Method

Overview

Considering that spatial-domain convolution excels at extracting local context information, while Fourier convolu-



(a) Overall architecture **(b) Frequency-Spatial Promoted and Physical Learning Block**

Figure 2: (a) The architecture of our proposed Joint Spatial and Fourier Convolutional Network (JSFC-Net) for single image dehazing. (b) The inner structure of Frequency-Spatial Promoted and Physical Learning (FS-PPL) Block.

tion captures long-range relationships and enhances high-frequency learning capabilities, we propose the Joint Spatial and Fourier Convolutional Network (JSFC-Net), as illustrated in Figure 2.

Given a hazy image, $\mathbf{I_{in}} \in \mathbb{R}^{H\times W\times 3}$, we first apply a 3x3 depth-wise separable convolutional layer (Howard et al. 2017) to extract shallow features. These shallow features are then sequentially processed through M Frequency-Spatial Promoted and Physical Learning (FS-PPL) blocks to capture high-level features. Finally, the reconstruction module projects high-level features back to original image size, producing the dehazed image.

It is noteworthy that each FS-PPL block is connected densely (Huang et al. 2017) (this is not shown in Figure 2). This dense connection enhances information flow throughout the network, aiding in the effective restoration of intricate details. To prevent information loss during the downsampling process, and because JSFC-Net directly acquires a global receptive field via FFT, we did not adopt the commonly used encoder-decoder architecture in image restoration networks.

Frequency-Spatial Promoted and Physical Learning Block

The FS-PPL block consists of the Frequency Promoted Learning Branch (FPLB) and the Spatial Physical Learning Branch (SPLB) in parallel, responsible for extracting global frequency features and local spatial features, respectively. Finally, the features from both domains are fused through the Receptive Field Selection Modul (RFSM).

Frequency Promoted Learning Branch Let $F_{in} \in \mathbb{R}^{C\times H\times W}$ denote the input features, which are initially transformed into frequency domain through FFT:

$$F_{fre} = \mathcal{F}(F_{in}), \tag{1}$$

where $\mathcal{F}(\cdot)$ is 2D FFT operator. $F_{fre} = \{R, I\}$ is the frequency features after FFT, where $\{R, I\} \in \mathbb{R}^{C\times H\times \lceil \frac{W}{2} \rceil}$ represents the real and imaginary parts, respectively. Here, $\lceil\cdot\rceil$ denotes the round-up operator. Please note that in accordance with the conjugate symmetry principle in Fourier transform theory, we only keep half of the spectrum to reduce computational cost. Previous methods (Li, You, and Robles-Kelly 2018; Zhang et al. 2022) only perform feature extraction on real part of Fourier spectrum, which loses a part of frequency information. Therefore, we concatenate the real and imaginary parts in the channel dimension. The concatenated tensors $F_{fre} \in \mathbb{R}^{2C\times H\times \lceil \frac{W}{2} \rceil}$ are all real numbers, modeling the comprehensive amplitude and phase information of frequency features. Subsequently, we apply 1x1 convolution and LeakyReLU for feature extraction, as depicted in Figure 2 (b).

Next, we utilize spatial domain features F_{spa}^{out} to promote and modulate F_{fre}, addressing the issue of insufficient robustness caused by the fact that each value in F_{fre} is globally sensitive. Specifically, we first utilize kernel-generating layers to produce a set of learnable convolution kernels with multiple scales, formulated as:

$$filter^k = \text{Softmax}\left(\text{BN}\left(\text{Conv1x1}\left(\text{GAP}(F_{spa}^{out})\right)\right)\right), \tag{2}$$

where $filter^k$ represents a learned convolution kernel with size of $k \times k$. We set k to be 3, 5, and 7, respectively, to perform feature modulation and refinement in a multi-scale manner. BN, Conv1x1, and GAP are Batch Normalization (Ioffe and Szegedy 2015), 1×1 convolution layer, and global average pooling, respectively. Then, we operate on F_{fre} using the $filter^k$, which is expressed as:

$$F_{fre}^k(i,j) = \sum_{m=-r}^{r} \sum_{n=-r}^{r} F_{fre}(i+m, j+n) \cdot filter^k(m,n), \tag{3}$$

where $r = \lfloor \frac{k}{2} \rfloor$, $\lfloor \cdot \rfloor$ denotes the round-down operator. Here, F_{fre}^k represents the result of promoting F_{fre} using learned convolution kernel $filter^k$ with size of $k \times k$. In this paper, we obtain F_{fre}^3, F_{fre}^5, and F_{fre}^7, respectively.

Drawing inspiration from classical Mixture-of-Experts (MoE) (Masoudnia and Ebrahimpour 2014), we employ a gated fusion sub-network, denoted by $\mathcal{G}$, to determine the contributions of each scale promoted frequency features. $\mathcal{G}$ is computationally inexpensive yet sufficiently expressive to make informative decisions, which can be expressed as:

$$\begin{aligned} (\sigma_1, \sigma_2, \sigma_3) &= \mathcal{G}(F_{fre}^3, F_{fre}^5, F_{fre}^7), \\ F_{fre}^* &= \sigma_1 * F_{fre}^3 + \sigma_2 * F_{fre}^5 + \sigma_3 * F_{fre}^7, \end{aligned} \tag{4}$$

where F_{fre}^* represents the final promoted frequency features. Finally, Inverse Fast Fourier Transform (IFFT) is applied to convert F_{fre}^* back to spatial domain, resulting in final output F_{fre}^{out} of FPLB.

Spatial Physical Learning Branch In terms of spatial domain feature learning, our objective is to integrate physical priors into the feature space, thereby promoting interpretability that aligns with the haze imaging equation (Narasimhan and Nayar 2002; Fattal 2008; Tan 2008). For the simplicity of network design and given the outstanding performance demonstrated by the physics-aware dual-branch Unit (PDU) proposed in (Zheng et al. 2023), we build SPLB on the basis of PDU.

PDU reformulates haze imaging equation (Narasimhan and Nayar 2002; Fattal 2008; Tan 2008) as:

$$F_{spa} = F_{in} \odot \tilde{t} + \tilde{A}(1 - \tilde{t}), \tag{5}$$

where F_{spa} represents physics-aware spatial domain output features and F_{in} denotes input features. $\tilde{t}$ and $\tilde{A}$ represent intermediate features corresponding to transmission map and atmospheric light, respectively. The expression for $\tilde{A}$ is:

$$\begin{aligned} \bar{A} = H(\sigma(\mathrm{Conv}_{1\times1}^{C}(\mathrm{LeakyReLU}(\\ \mathrm{Conv}_{1\times1}^{\frac{C}{8}}(\mathrm{GAP}(\mathrm{F}_{in})))))), \end{aligned} \tag{6}$$

where $\sigma(\cdot)$ is the Sigmoid function, $H(\cdot)$ denotes a replication operation, and $\mathrm{Conv}_{1\times1}^{C}$ represents a convolutional layer with output channels C and kernel size 1. The expression for $\tilde{t}$ is given by:

$$\begin{aligned} \tilde{t} = \sigma(\mathrm{Conv3x3}^{C}(\mathrm{LeakyReLU}(\\ \mathrm{Conv}_{3\times3}^{\frac{C}{8}}(\mathrm{Conv}_{3\times3}^{C}(F_{in}))))), \end{aligned} \tag{7}$$

SPLB consists of a stack of N blocks. Finally, we obtain the output F_{spa}^{out} of SPLB.

Receptive Field Selection Module We believe that due to differences in local details and global background, as well as differences in the scale of image features, varying pixels require varying receptive field. For example, textures and edges containing detailed information need a smaller receptive field (*i.e.*, F_{spa}^{out}) to capture features more precisely, while background or large haze regions require a larger receptive field (*i.e.*, F_{fre}^{out}) for effective processing.

Following this insight, we introduce RFSM to dynamically explore the optimal receptive field for each pixel to better fuse spatial and frequency domain features. Specifically, we generate the receptive field maps as follows:

$$\begin{aligned} (M_{short}, M_{long}) = &\mathrm{Softmax}(\mathrm{Conv}_{1\times1}(\\ &\mathrm{DWConv}([F_{in}, F_{spa}^{out}, F_{fre}^{out}]))), \end{aligned} \tag{8}$$

where $[\cdot]$ denotes concatenation, DWConv represents depth-wise separable convolutional layer (Howard et al. 2017), and $\mathrm{Conv}_{1\times1}$ is a convolutional layer with an output channel of 2. Given that $F_{in}, F_{spa}^{out}, F_{fre}^{out} \in \mathbb{R}^{C\times H\times W}$ and $M_{short}, M_{long} \in \mathbb{R}^{1\times H\times W}$, the output of FS-PPL block is:

$$F_{out} = M_{short} \times F_{spa}^{out} + M_{long} \times F_{fre}^{out}. \tag{9}$$

Loss Function

We adopt $L1$ loss $\mathcal{L}_1$ and perceptual loss (Johnson, Alahi, and Fei-Fei 2016) $\mathcal{L}_P$ to optimize the proposed JSFC-Net in spatial domain. The former prevents the suppression of high-frequency details caused by MSE loss (Gondal, Schölkopf, and Hirsch 2018), while the latter ensures that dehazed images conform more closely to human visual perception.

However, existing deep dehazing networks only focus on minimizing pixel gaps between the restored image and the label. This minimization ignores the frequency gaps, resulting in inefficient recovery for hard frequency (*e.g.*, high-frequency details). To tackle this challenge, we introduce frequency loss to directly optimize the network in the frequency domain, which is formulated as follows:

$$\mathcal{L}_{fre} = \left\| \mathcal{F}(I_{gt}) - \mathcal{F}(Net(I_{haze})) \right\|_1, \tag{10}$$

where I_{gt} and I_{haze} are the ground truth image and hazy input image, respectively; Net is our JSFC-Net. The overall loss function is written as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_P + \lambda_3 \mathcal{L}_{fre}, \tag{11}$$

where λ_1, λ_2, and λ_3 represent the weight parameters.

Experiments

Experimental Settings

Implementation Details The JSFC-Net is implemented using the PyTorch framework on an Intel Gold 6252 CPU and NVIDIA A100 GPUs. We use the Adam (Kingma and Ba 2014) optimizer with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.99$) and a cosine annealing strategy (Loshchilov and Hutter 2016) to train JSFC-Net. The initial learning rate is set to 1×10^{-4} and gradually decreases to 5×10^{-6}.

Methods	Publication	SOTS-Indoor		SOTS-Outdoor		NH-Haze		O-Haze		#Param	Time
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	(M)	(ms)
DCP (He, Sun, and Tang 2010)	TPAMI-10	16.62	0.818	19.13	0.815	12.11	0.452	16.78	0.653	-	1.70
DehazeNet (Cai et al. 2016)	TIP-16	19.82	0.821	24.75	0.927	12.38	0.455	17.57	0.770	0.01	0.63
GridDehazeNet (Liu et al. 2019)	ICCV-19	32.16	0.984	30.86	0.982	18.19	0.608	23.48	0.726	0.96	14.95
MSBDN (Dong et al. 2020)	CVPR-20	33.79	0.984	33.48	0.982	19.35	0.640	24.36	0.749	31.35	29.77
FFA-Net (Qin et al. 2020)	AAAI-20	36.39	0.989	33.57	0.984	19.27	0.637	22.12	0.770	4.46	35.98
AECR-Net (Wu et al. 2021)	CVPR-21	37.17	0.990	-	-	19.46	0.641	23.21	0.749	2.61	6.08
MAXIM-2S (Tu et al. 2022)	CVPR-22	38.11	0.991	34.19	0.985	-	-	-	-	14.10	178.55
DeHamer (Guo et al. 2022)	CVPR-22	36.63	0.988	35.18	0.986	20.66	0.680	25.11	0.777	132.50	18.74
DehazeFormer-M (Song et al. 2023)	TIP-23	38.46	0.994	34.29	0.983	20.60	0.670	<u>25.48</u>	0.765	4.63	35.85
MITNet (Shen et al. 2023)	MM-23	40.23	0.992	35.18	0.988	<u>21.26</u>	0.712	-	-	2.73	14.15
FocalNet (Cui et al. 2023a)	ICCV-23	<u>40.82</u>	**0.996**	<u>37.71</u>	**0.995**	20.43	<u>0.790</u>	-	-	3.74	10.81
MB-TaylorFormer-B (Qiu et al. 2023)	ICCV-23	40.71	0.992	37.42	0.989	-	-	25.05	<u>0.788</u>	2.68	177.74
DEA-Net (Chen, He, and Lu 2024)	TIP-24	40.20	0.993	36.03	0.989	-	-	-	-	3.65	32.35
JSFC-Net	-	**41.32**	**0.996**	**37.84**	<u>0.994</u>	**21.42**	**0.799**	**25.72**	**0.804**	3.73	22.38

Table 1: Quantitative evaluation on RESIDE (Li et al. 2018), NH-HAZE (Ancuti, Ancuti, and Timofte 2020) and O-Haze (Ancuti et al. 2018) datasets. The best results and the second best results are in **bold** and <u>underline</u>, respectively.

Setting	SOTS-Indoor		NH-Haze		O-Haze	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
(1) − Frequency Promoted Learning Branch	38.68	0.992	20.67	0.725	24.89	0.762
(2) − Spatial Physical Learning Branch	21.44	0.822	14.28	0.574	18.85	0.687
(3) → Randomly initialized kernels	40.58	0.994	20.99	0.774	25.33	0.778
(4) → Single scale learnable kernels	41.21	0.995	21.18	0.789	25.47	0.786
(5) − Mixture-of-Experts	41.17	0.995	21.24	0.793	25.62	0.791
(6) − Receptive Field Selection Modul	40.99	0.994	21.03	0.781	25.36	0.779
JSFC-Net	**41.32**	**0.996**	**21.42**	**0.799**	**25.72**	**0.804**

Table 2: Quantitative results of the ablation experiments on the proposed key components. The best results are in **bold**.

Datasets For a comprehensive comparison, we evaluate the proposed JSFC-Net on synthetic haze datasets (*i.e.*, RESIDE (Li et al. 2018)), generated haze datasets (*i.e.*, NH-HAZE (Ancuti, Ancuti, and Timofte 2020) and O-Haze (Ancuti et al. 2018) datasets) and real-world hazy images. In RESIDE (Li et al. 2018), two subsets, the Indoor Training Set (ITS) and the Outdoor Training Set (OTS), are selected for training, consisting of 13,990 pairs and 313,950 pairs of images, respectively. Models are evaluated on the Synthetic Objective Testing Set (SOTS) subset. The NH-HAZE (Ancuti, Ancuti, and Timofte 2020) and O-Haze (Ancuti et al. 2018) datasets contain 55 and 45 images, respectively. We select 5 images from each dataset as test sets and the remaining images as training sets.

Comparison to State-of-the-Art Methods

Results on Synthetic Haze Dataset The quantitative results on the widely used synthetic dataset RESIDE are shown in Table 1. Our JSFC-Net achieves the highest PSNR on both indoor and outdoor test sets. Specifically, compared to FocalNet with the second best performance, our JSFC-Net achieves a PSNR improvement of 0.50 dB in indoor scenes and 0.12 dB in outdoor scenes.

Figure 3 shows visual comparisons on the RESIDE SOTS indoor dataset. The red box highlights low-frequency structures, while the green box highlights high-frequency texture details. Although MITNet (Shen et al. 2023) and DEA-Net (Chen, He, and Lu 2024) achieve competitive quantitative results, they still struggle to restore high-frequency details effectively. For instance, in Figure 3, both MITNet (Shen et al. 2023) and DEA-Net (Chen, He, and Lu 2024) fail to adequately restore the ear area of the man, resulting in an incomplete facial appearance. Our JSFC-Net produces the most natural results, closely resembling the patterns observed in the ground truth for both low-frequency structures and high-frequency details.

Results on Generated Haze Dataset We also compare JSFC-Net with SOTA methods on NH-HAZE and O-Haze datasets, which use a professional haze generator to simulate real hazy conditions. As shown in Table 1, JSFC-Net consistently outperforms other methods. Specifically, compared to the second-best method, JSFC-Net achieves a performance gain of 0.16 dB in PSNR and 0.009 in SSIM for NH-HAZE, and 0.24 dB in PSNR and 0.016 in SSIM for O-Haze.

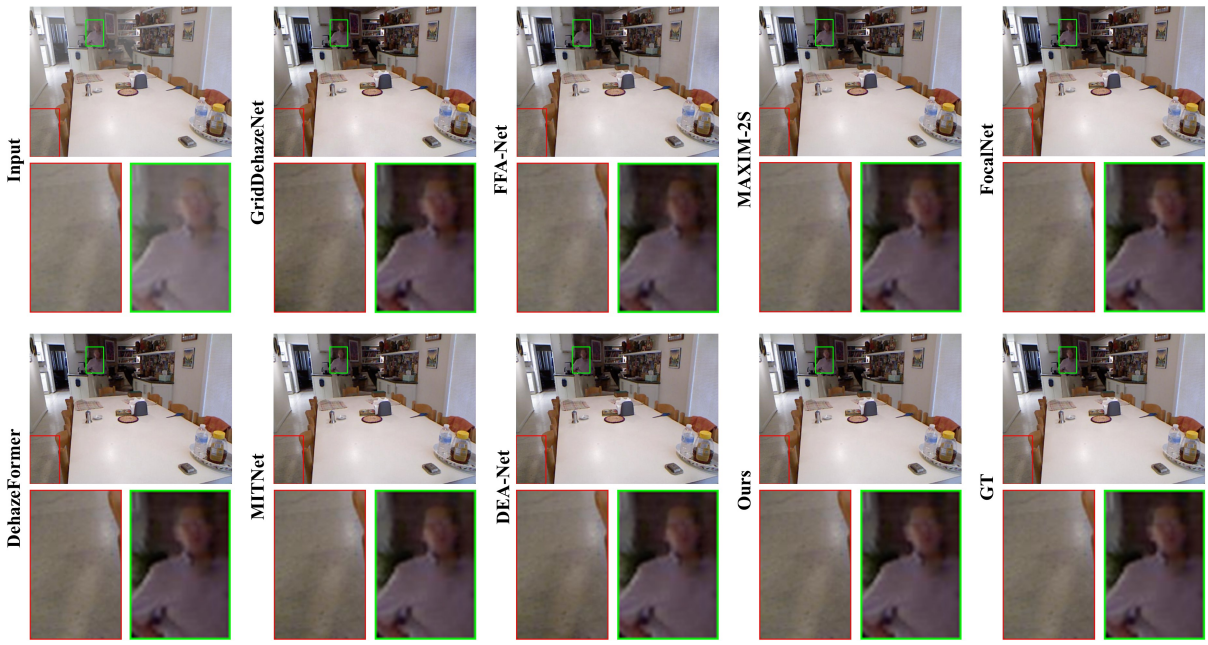


Figure 3: Visual comparisons on the RESIDE SOTS indoor dataset. Zoom in for the best view.

Results on Real-World Hazy Images To further assess the generalization capability of JSFC-Net, trained on synthetic haze datasets, to real-world hazy scenes, we present the dehazing results on natural hazy images, as illustrated in Figure 4. To ensure that the comparison methods are well-trained, we use the pre-trained weights provided by the authors on the RESIDE Outdoor dataset. DCP and GridDehazeNet produce impressive results, effectively restoring the details of buildings (as seen in the green boxes), while other SOTA methods exhibit haze residuals. However, both GridDehazeNet and DCP show haze residuals and unrealistic details in the tree areas, respectively. In contrast, JSFC-Net produces the most favorable results in both local textures and overall visual quality. This demonstrates that JSFC-Net not only achieves SOTA performance on synthetic datasets but also generalizes well to real-world scenes, offering an effective solution for real-world image dehazing.

Ablation Study

We conduct ablation studies on the RESIDE indoor dataset to evaluate the effectiveness of the core components. The hyperparameters are adjusted to ensure that all variant models have nearly the same number of parameters as JSFC-Net.

Joint Learning in Spatial and Frequency Domain To evaluate whether joint learning of spatial and frequency domain features can enhance dehazing model performance, we conducted ablation experiments. Specifically, based on the original JSFC-Net, we construct two variants: model (1), which excludes the FPLB component, and model (2), which excludes the SPLB component. Model (2) consists solely of convolutional layers operating on the frequency spectrum, without spatial domain feature modulation and MoE module. Both model (1) and model (2) also omit the RFSM.

As demonstrated in Table 2, relying solely on either the spatial or frequency domain fails to yield satisfactory results, particularly in the latter case. However, when combining the two domains, an improvement of 2.64 dB in PSNR and 0.004 in SSIM is achieved compared to the second-best performance. These results also suggest that frequency features are best utilized as a supplement to spatial features and cannot be relied upon alone for image restoration.

Core Components in Frequency-Spatial Promoted and Physical Learning Block Ablation experiments are conducted to evaluate the effectiveness of each component in the proposed FS-PPL block, including the modulation of frequency features using spatial features, multi-scale learnable kernels, the MoE module, and RFSM.

To validate the significance of promoting and modulating frequency features with spatial features, we first replace the learnable kernels in the original JSFC-Net with three randomly initialized convolution kernels of sizes 3x3, 5x5, and 7x7, resulting in model (3). Next, we evaluate the importance of multi-scale kernels by substituting the original multi-scale kernels (3x3, 5x5, and 7x7) with kernels of the same size. For a fair comparison, we use four 5x5 learnable kernels based on spatial domain features, creating model (4). Additionally, we replace the MoE module in JSFC-Net with element-wise addition to form model (5). Finally, we examine the impact of RFSM by constructing model (6) using element-wise addition. The quantitative results are presented in Table 2. All of these strategies lead to performance improvements across multiple indicators on various datasets.

Universal Frequency Loss

We incorporate frequency loss into the training process of SOTA methods to assess its general applicability. As shown

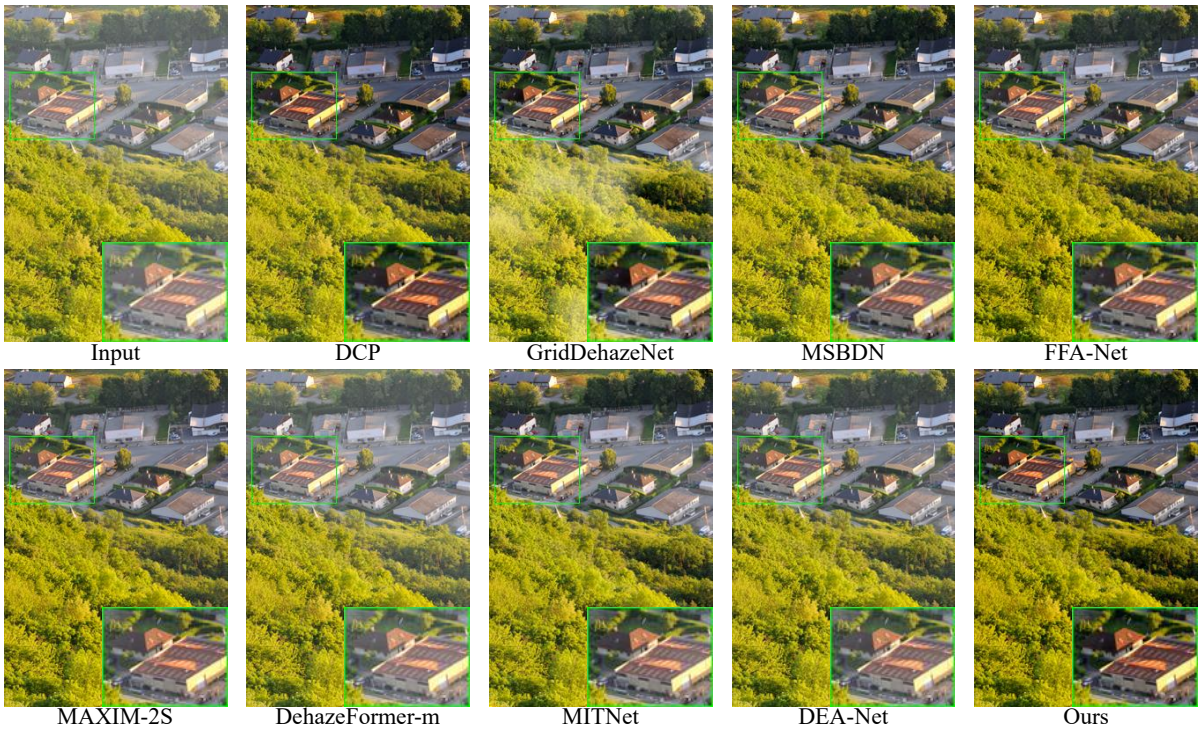


Figure 4: Visual comparisons on real-world hazy image. Zoom in for the best view.

Method (+ Frequency Loss)	PSNR	SSIM
AOD-Net (Li et al. 2017)	↑1.94	↑0.0766
GridDehazeNet (Liu et al. 2019)	↑0.47	↑0.0015
MSBDN (Dong et al. 2020)	↑0.09	↑0.0011
PCFAN (Zhang et al. 2020)	↑0.22	↑0.0005
FFA-Net (Qin et al. 2020)	↑1.19	↑0.0032
AECR-Net (Wu et al. 2021)	↑0.30	↑0.0019
DeHamer (Guo et al. 2022)	↑0.19	↑0.0006
DehazeFormer-M (Song et al. 2023)	↑0.08	↑0.0003
DEA-Net (Chen, He, and Lu 2024)	↑0.11	↑0.0005

Table 3: Quantitative results of applying frequency loss into SOTA methods. ↑ denotes performance gains

in Table 3, the inclusion of frequency loss consistently improves the performance of these SOTA methods. This demonstrates that emphasizing discrepancies between the restored image and ground truth in the frequency domain enhances the effectiveness of supervised dehazing algorithms, without adding extra parameters for inference.

Deployability

To verify model's deployability, we provide the inference time at a resolution of 256x256 in Table 1. Additionally, the performance and efficiency trade-offs are compared in Figure 1. Our JSFC-Net achieves the best trade-offs, with a frame per second rate of 45, meeting real-time requirements.

Conclusion

In this paper, we design JSFC-Net for image dehazing. Unlike previous methods that relied on deeper networks or encoder-decoder architectures to enlarge receptive fields, JSFC-Net utilizes Fourier transform to achieve a global receptive field with lower overhead. Another benefit is that extracting features in the frequency domain addresses the common challenge of high-frequency learning in existing methods. The core building block of JSFC-Net is the Frequency-Spatial Promoted and Physical Learning Block, which significantly enhances the model's expressive power by leveraging dual-domain features, namely physical spatial and promoted frequency domain features. We design the Receptive Field Selection Modul to determine the optimal receptive field for each pixel, thereby enhancing the integration of spatial and frequency domain. Finally, we introduce the frequency loss, which reduces the disparities between dehazed and clear images in the frequency domain, providing a valuable complement to existing spatial losses. We conduct extensive experiments on three benchmark datasets and real-world hazy images, verifying the superior performance of JSFC-Net compared to SOTA methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62475006, and in part by the Beijing Municipal Natural Science Foundation under Grant L242106.

References

Ancuti, C.; Ancuti, C. O.; Timofte, R.; and Vleeschouwer, C. D. 2018. I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images. In *Int. Conf. Adv. Conce. Intell. Vis. Syst.*, 620–631. Springer.

Ancuti, C. O.; Ancuti, C.; and Timofte, R. 2020. NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proc. Conf. Comput. Vis. Pattern Recognit. workshops*, 444–445.

Cai, B.; Xu, X.; Jia, K.; Qing, C.; and Tao, D. 2016. Dehazenet: An end-to-end system for single image haze removal. *IEEE Trans. Image Process.*, 25(11): 5187–5198.

Chen, Z.; He, Z.; and Lu, Z.-M. 2024. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*.

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast fourier convolution. *Adv. Neural Inf. Process. Syst.*, 33: 4479–4488.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2023a. Focal network for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13001–13011.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2023b. Image restoration via frequency selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; and Yang, M.-H. 2020. Multi-scale boosted dehazing network with dense feature fusion. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2157–2167.

Fattal, R. 2008. Single image dehazing. *ACM Trans. Graph. (TOG)*, 27(3): 1–9.

Fattal, R. 2014. Dehazing using color-lines. *ACM Trans. Graph. (TOG)*, 34(1): 1–14.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. In *Int. Conf. Mach. Learn.*, 3247–3258. PMLR.

Gondal, M. W.; Schölkopf, B.; and Hirsch, M. 2018. The unreasonable effectiveness of texture transfer for single image super-resolution. In *Eur. Conf. Comput. Vis.*, 80–97. Springer.

Guo, C.-L.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; and 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*, 5812–5820.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12): 2341–2353.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. pmlr.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021. Focal frequency loss for image reconstruction and synthesis. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 13919–13929.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Katznelson, Y. 2004. *An introduction to harmonic analysis*. Cambridge University Press.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, L.; Dong, J.; Ge, J.; Li, M.; and Pan, J. 2023. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5886–5895.

Kulkarni, A.; Phutke, S. S.; and Murala, S. 2022. Unified transformer network for multi-weather image restoration. In *European Conference on Computer Vision*, 344–360. Springer.

Li, B.; Peng, X.; Wang, Z.; Xu, J.; and Feng, D. 2017. Aod-net: All-in-one dehazing network. In *Proc. IEEE. Int. Conf. Comput. Vis.*, 4770–4778.

Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2018. Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.*, 28(1): 492–505.

Li, C.; Zhou, H.; Liu, Y.; Yang, C.; Xie, Y.; Li, Z.; and Zhu, L. 2023. Detection-friendly dehazing: Object detection in real-world hazy scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8284–8295.

Li, J.; You, S.; and Robles-Kelly, A. 2018. A frequency domain neural network for fast image super-resolution. In *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Liu, X.; Ma, Y.; Shi, Z.; and Chen, J. 2019. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 7314–7323.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Masoudnia, S.; and Ebrahimpour, R. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42: 275–293.

McCartney, E. J. 1976. Optics of the atmosphere: scattering by molecules and particles. *New York*.

Miao, Y.; Deng, J.; and Han, J. 2024. WaveFace: Authentic Face Restoration with Efficient Frequency Recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6583–6592.

Narasimhan, S. G.; and Nayar, S. K. 2002. Vision and the atmosphere. *Int. J. Comput. Vis.*, 48(3): 233–254.

Nayar, S. K.; and Narasimhan, S. G. 1999. Vision in bad weather. In *Proc. IEEE Int. Conf. Comput. Vis.*, volume 2, 820–827. IEEE.

Pan, J.; Sun, D.; Zhang, J.; Tang, J.; Yang, J.; Tai, Y.-W.; and Yang, M.-H. 2022. Dual Convolutional Neural Networks for Low-Level Vision. *Int. J. Comput. Vis.*, 130(6): 1440–1458.

Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; and Jia, H. 2020. FFA-Net: Feature fusion attention network for single image dehazing. In *Proc. AAAI Conf. Artif. Intell.*, volume 34, 11908–11915.

Qiu, Y.; Zhang, K.; Wang, C.; Luo, W.; Li, H.; and 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*, 12802–12813.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global filter networks for image classification. *Adv. Neural Inf. Process. Syst.*, 34: 980–993.

Ren, W.; Ma, L.; Zhang, J.; Pan, J.; Cao, X.; Liu, W.; and Yang, M.-H. 2018a. Gated fusion network for single image dehazing. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, 3253–3261.

Ren, W.; Pan, J.; Zhang, H.; Cao, X.; and Yang, M.-H. 2020. Single image dehazing via multi-scale convolutional neural networks with holistic edges. *Int. J. Comput. Vis.*, 128(1): 240–259.

Ren, W.; Zhang, J.; Xu, X.; Ma, L.; Cao, X.; Meng, G.; and Liu, W. 2018b. Deep video dehazing with semantic segmentation. *IEEE transactions on image processing*, 28(4): 1895–1908.

Rippel, O.; Snoek, J.; and Adams, R. P. 2015. Spectral representations for convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 28.

Shen, H.; Zhao, Z.-Q.; Zhang, Y.; and Zhang, Z. 2023. Mutual information-driven triple interaction network for efficient image dehazing. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7–16.

Song, Y.; He, Z.; Qian, H.; and Du, X. 2023. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32: 1927–1941.

Tan, R. T. 2008. Visibility in bad weather from a single image. In *IEEE Conf. Comput. Vis. Pattern. Recognit.*, 1–8. IEEE.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5769–5780.

Wang, Y.; Xiong, J.; Yan, X.; and Wei, M. 2023. USCFormer: unified transformer with semantically contrastive learning for image dehazing. *IEEE Transactions on Intelligent Transportation Systems*.

Wang, Y.; Xu, C.; You, S.; Tao, D.; and Xu, C. 2016. Cnnpack: Packing convolutional neural networks in the frequency domain. *Adv. Neural Inf. Process. Syst.*, 29.

Wei, Y.; Gu, S.; Li, Y.; Timofte, R.; Jin, L.; and Song, H. 2021. Unsupervised real-world image super resolution via domain-distance aware training. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 13385–13394.

Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; and Ma, L. 2021. Contrastive learning for compact single image dehazing. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 10551–10560.

Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.-K.; and Ren, F. 2020. Learning in the frequency domain. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 1740–1749.

Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; and Ha, J.-W. 2019. Photorealistic style transfer via wavelet transforms. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 9036–9045.

Yu, H.; Zheng, N.; Zhou, M.; Huang, J.; Xiao, Z.; and Zhao, F. 2022. Frequency and spatial dual guidance for image dehazing. In *European Conference on Computer Vision*, 181–198. Springer.

Zhang, D.; Huang, F.; Liu, S.; Wang, X.; and Jin, Z. 2022. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*.

Zhang, H.; and Patel, V. M. 2018. Densely connected pyramid dehazing network. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, 3194–3203.

Zhang, X.; Wang, T.; Wang, J.; Tang, G.; and Zhao, L. 2020. Pyramid channel-based feature attention network for image dehazing. *Comput. Vis. Image Understanding*, 197: 103003.

Zhang, X.; Xie, F.; Ding, H.; Yan, S.; and Shi, Z. 2024. Proxy and Cross-Stripes Integration Transformer for Remote Sensing Image Dehazing. *IEEE Transactions on Geoscience and Remote Sensing*.

Zheng, Y.; Zhan, J.; He, S.; Dong, J.; and 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*, 5785–5794.

Zhou, M.; Huang, J.; Yan, K.; Yu, H.; Fu, X.; Liu, A.; Wei, X.; and Zhao, F. 2022. Spatial-frequency domain information integration for pan-sharpening. In *European conference on computer vision*, 274–291. Springer.

Zhu, Q.; Mai, J.; and Shao, L. 2015. A fast single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Process.*, 24(11): 3522–3533.