# When Fast Fourier Transform Meets Transformer for Image Restoration

Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng*

School of Astronautics, Beihang University, Beijing, China
ydeng@buaa.edu.cn

**Abstract.** Natural images can suffer from various degradation phenomena caused by adverse atmospheric conditions or unique degradation mechanism. Such diversity makes it challenging to design a universal framework for kinds of restoration tasks. Instead of exploring the commonality across different degradation phenomena, existing image restoration methods focus on the modification of network architecture under limited restoration priors. In this work, we first review various degradation phenomena from a frequency perspective as prior. Based on this, we propose an efficient image restoration framework, dubbed SFHformer, which incorporates the Fast Fourier Transform mechanism into Transformer architecture. Specifically, we design a dual domain hybrid structure for multi-scale receptive fields modeling, in which the spatial domain and the frequency domain focuses on local modeling and global modeling, respectively. Moreover, we design unique positional coding and frequency dynamic convolution for each frequency component to extract rich frequency-domain features. Extensive experiments on thirty-one restoration datasets for a range of ten restoration tasks such as deraining, dehazing, deblurring, desnowing, denoising, super-resolution and underwater/low-light enhancement, demonstrate that our SFHformer surpasses the state-of-the-art approaches and achieves a favorable trade-off between performance, parameter size and computational cost. The code is available at: https://github.com/deng-ai-lab/SFHformer.

**Keywords:** Image Restoration · Frequency Feature · Deep Learning

## 1 Introduction

Natural images can suffer from different degradation processes (see Fig.1a), resulting from various adverse weather conditions (haze [6], snow [15], rain [82] and water [40]), lens defocus (defocus blur [2]), relative motion (motion blur [55]) and underexposure (low-light [84]), which greatly affects kinds of downstream visual tasks [68,105]. As a result, image restoration as a solution, aiming to reconstruct the low-quality degraded images into the high-quality clear ones, has been

* Corresponding author

widely discussed in recent decades. However, due to the ill-posed intrinsic properties of image restoration, it is challenging to design an efficient image restoration model for effectively handling various degradation processes. So far, existing restoration approaches can be well divided into two categories: traditional prior-based approaches [7, 28, 95] and deep data-driven approaches [37, 48, 98]. Prior-based models regard image restoration as optimization problem and utilize physical assumptions to constrain the solution space, but may fail in challenging scenarios with complicated structures. On the other hand, data-driven models, especially CNNs, adopt the end-to-end manner to accomplish image restoration and obtain remarkable performance. More recently, transformer-like global modeling models [74,83] have emerged and achieved the most state-of-the-art(SOTA) performance on kinds of image restoration tasks.

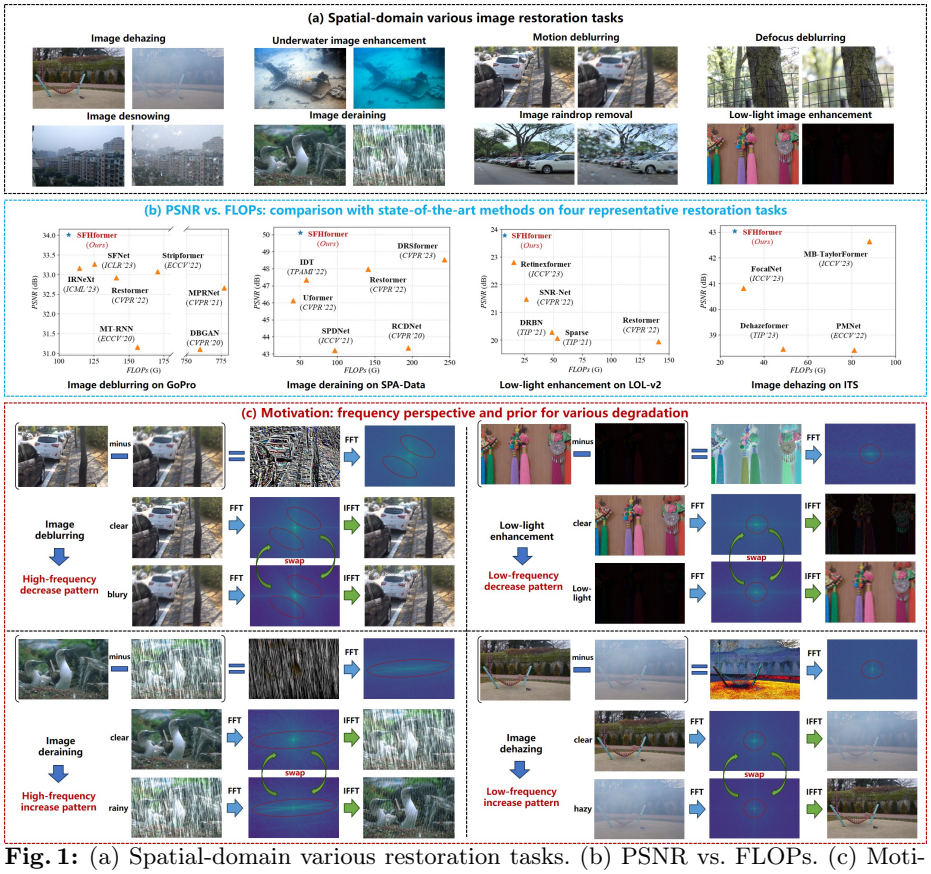


**Fig. 1:** (a) Spatial-domain various restoration tasks. (b) PSNR vs. FLOPs. (c) Motivation: frequency-domain perspective and prior for various degradation phenomena.

Although recent transformer-like global modeling approaches have made substantial progress for image restoration, they still suffer from several shortcom-

ings. First, most existing global modeling restoration methods hardly consider the common prior properties of various degradation processes in network design. With limited restoration priors, these approaches may encounter difficulties on some specific restoration tasks. As shown in Fig.1b, the SOTA method Restormer [98] obtains comparable performance on deblurring and deraining tasks, but fails on low-light enhancement task. Second, existing global modeling mechanisms, such as self-attention [49], require significant computational resources. While numerous methods have been optimized for self-attention mechanisms [65, 98], they suffer from performance decrease and fail to achieve a favorable balance between performance, parameter size and computational cost.

Driven by above two aforementioned problems, we first revisit various degradation phenomena from a frequency perspective and observe that frequency-domain features inherently embody discernible restoration priors, which can be properly leveraged in network design. Next, we elucidate the rationale behind frequency perspective for various degradation phenomena from the following three aspects (see Fig.1c). **(1)Distinctive and compact frequency representation of various degradation processes.** The first row of four sub-regions in Fig.1c displays the residual of the clear and degraded images and its frequency version via Fast Fourier Transform(FFT). In deraining and deblurring, the degradation is mainly manifested in high-frequency domain, where rain streak manifests as increase pattern, while blurring manifests as decrease pattern. On the other hand, in dehazing and low-light, the degradation is mainly manifested in low-frequency domain, where haze manifests as increase pattern, while low-light manifests as decrease pattern. In addition, the frequency-domain degradation residual exhibits more compact representation than spatial-domain ones, which imply that modeling degradation from a frequency-domain perspective offers more efficiency. **(2)Strong frequency prior for restoration.** In the two subsequent rows in Fig.1c, we swap the compact degradation representation region(see red circle) of clear and degraded images based on discovery of **(1)**. Consequently, the initially clear image transform into its degraded counterpart, and conversely, the degraded image exhibits clear version. Therefore, the FFT operator possesses the capability to disentangle ground-truth and degradation from a frequency-domain perspective to a significant extent, thereby serving as a strong prior for various restoration tasks. **(3)Efficient global receptive field properties.** It is worth mentioning that the FFT operation inherently possesses the global modeling property, that is, changing a certain frequency component in the frequency-domain will affect the pixel in the entire spatial-domain. Besides, FFT offers $O(NlogN)$ computational complexity superior than self-attention, which has $O(N^2)$ computational complexity. Building upon the aforementioned considerations and drawing inspiration from [97], we incorporate FFT operator into the token mixer module to facilitate efficient global modeling, replacing vanilla self-attention mechanisms. This enables us to effectively model various degradation processes by extracting intricate frequency features.

In this work, we propose a transformer-like image restoration backbone, dubbed SFHformer, for restoring various degradation process under frequency

prior. Specifically, SFHformer is comprised of two essential modules: Local-Global Perception Mixer(LGPM) and Multi-kernel ConvFFN(MCFN). Inside LGPM, we design a novel Spatial-Frequency domain Hybrid structure to replace the vanilla self-attention for less computational complexity, in which the spatial-domain branch focuses on extracting local features and the frequency-domain branch concentrates on capturing global relationships. Moreover, to better distinguish each frequency component in frequency-domain branch, we introduce the implicit position encoding to dynamically assign unique identification and, inspired by [39, 89], we design the frequency dynamic convolution to capture customized frequency features. Towards MCFN, we bring in the multi-scale representation learning [76] to aggregate local and global features through multi-kernel receptive fields. The multi-kernel mechanism in MCFN can model context interactions from different perspectives and better preserve details. To validate the proposed model's effectiveness, we conduct comprehensive experiments and demonstrate state-of-the-art performance of our SFHformer on 31 benchmark datasets for a range of image restoration tasks, including image deraining, image dehazing, image desnowing, image raindrop removal, motion deblurring, single-image defocus deblurring, image denoising, image super-resolution, underwater image enhancement and low-light image enhancement.

We summarize the main contributions as three-folds:

- We provide a frequency-domain perspective and prior for reviewing various degradation process, which can be utilized as new insights for the design of image restoration networks.
- We propose SFHformer as a novel efficient backbone for kinds of image restoration tasks, which extracts the local-global features through spatial-frequency hybrid domain and introduces multi-scale representation learning to aggregate them.
- We show SFHformer surpasses existing state-of-the-art methods and maintains a favorable balance between performance, parameter size and computational cost on 31 image restoration benchmarks within 10 different restoration tasks.

## 2 Related Works

**Image Restoration.** All along, image restoration, aiming to restore a degraded image to a clear counterpart, has been discussed for a long time. To tackle this problem, existing image restoration methods can be divided into two categories: the prior-based methods [7, 28, 95] and the data-driven methods [11, 56, 85, 98]. Most prior-based approaches employ novel physical assumptions to constrain the solution space. For instance, He et al. [28] proposes the dark channel prior for image dehazing. Yi et al. [95] proposes the residue channel prior for image deraining. These prior-based methods show nice statistical properties in specific scenes, but can easily fail in real-world images where the physical assumptions do not hold.

Recently, data-driven image restoration methods, particularly CNNs, have been proposed to overcome shortcomings of traditional prior-based methods. Benefited from novel architecture design (such as residual [29] and dense block [46]) or advanced mechanism (such as attention [64] and multi-stage mechanism [22]), these CNN-based methods achieve impressive performance in a end-to-end manner. More recently, Vision Transformer [49], which is first developed for sequence processing in natural language tasks, has gained popularity in the low-level vision community. Different from CNNs, Vision Transformer captures long-range dependencies from a global modeling perspective. Various transformer-based algorithms [44,65,98] have validated the importance of global modeling for image restoration and achieved superior performance in image dehazing [38], image deraining [91], motion deblurring [55] and low-light image enhancement [84]. However, these transformer architectures, which adopt self-attention to realize global modeling, suffer from huge computation cost. Therefore, in our work, a frequency-domain operation is carefully proposed as an alternative strategy to conduct efficient global modeling.

**Frequency Features.** Many algorithms [17, 67, 75, 107] have been developed to address various computer vision problems through aggregating features in frequency domain. Early, frequency features are used for high-level vision tasks to represent non-local relationships. For instance, Chi et al. [17] propose the Fast Fourier Convolution to directly extract features in frequency domain and Rao et al. [67] propose the global filter network to control features of different frequency component for image classification. Recently, with more research [11,16,77,83] verifying that global modeling is also effective for low-level vision tasks, frequency-domain features have shown great potential in image restoration tasks. For instance, Suvorov et al. [75] introduce the Fast Fourier Convolution for image inpainting and Cui et al. [22] utilizes frequency selection mechanism to choose the most informative frequency to recover. Different from existing frequency-domain restoration methods, our work highlights in following aspects: **(1)Deeper motivation from frequency representation and degradation process.** Many existing models incorporate frequency based primarily on its high global modeling efficiency, often without comprehensive analysis for restoration tasks. In our research, we initiate a detailed examination of frequency representations across various degradation processes and demonstrate that frequency effectively serves as a robust prior for restoration. Through detailed visualizations and analyses, we interpret ***"why frequency features are significantly effective in aiding restoration?"***, which lays a foundational insight for future research. **(2)More straightforward implementation for frequency global modeling.** Current models achieve frequency global modeling through various methodologies: some serve as a complement to self-attention mechanism [60], some incorporate *convolution theorem* into self-attention [35], while others utilize frequency filter via gating [41,67]. Inspired by MetaFormer, we empirically substantiate that frequency operation can efficiently replace self-attention and introduce 'dynamic frequency convolution' to directly extract frequency features by learnable parameters, rather than filtering.

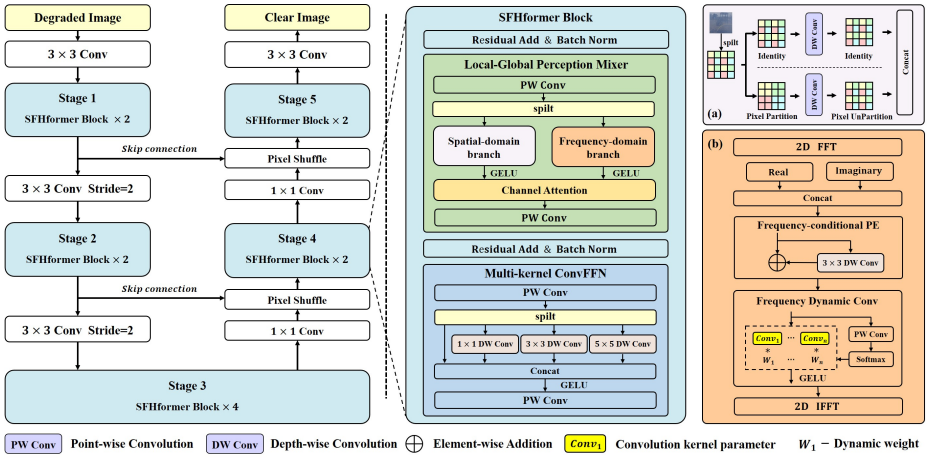


**Fig. 2:** Overview of the proposed SFHformer. (a) details of spatial-domain branch and (b) details of frequency-domain branch.

## 3 Methods

The proposed SFHformer utilizes a hierarchical encoder-decoder structure of five stages: a two-scale encoder (stage-1 and stage-2), a bottleneck (stage-3) and a two-scale decoder (stage-4 and stage-5). In the remaining part of this section, we will first present the overall architecture of our SFHformer (see Fig.2) and then describe the core components of the proposed SFHformer block: (a) Local-Global Perception Mixer(LGPM) and (b) Multi-kernel ConvFFN(MCFN).

### 3.1 Overall Architecture

As illustrated in Fig.2, given a degraded image $x \in \mathbb{R}^{H\times W\times 3}$, SFHformer first applies a $3 \times 3$ convolution to extract low-level features $F_0 \in \mathbb{R}^{H\times W\times C}$;where $H \times W$ denotes the spatial dimension and $C$ is the number of channels. Next, the shallow features $F_0$ are successively passed through a 5-stage hierarchical encoder-decoder structure. Between each stage, the spatial resolution is gradually reduced through a downsampling operation (3×3, stride 2 convolution layer) in the encoder and is then increased through a upsampling operation (pixel-shuffle) in the decoder. Inside each stage, the input features $F_0$ is passed through $N$ SFHformer blocks to extract multi-scale latent features $\{F_l^i, i = 1, ..., 5\} \in \{\mathbb{R}^{\frac{H}{n}\times\frac{W}{n}\times nC}, n = 1, 2, 4\}$. Then, to maintain the structural and textural features for restoration, the low-level latent features $F_l^i, i = 1, 2$ are concatenated with the high-level latent features $F_l^i, i = 3, 4$ via skip connection. Finally, a $3 \times 3$ convolution layer is applied to the final latent features $F_l^5$, generating the residual image $s \in \mathbb{R}^{H\times W\times 3}$ for obtaining the clear image $\hat{x}$ via $\hat{x} = x+s$. In the next two subsections, we will present the configuration of the two fundamental modules: LGPM and MCFN.

### 3.2 Local-Global Perception Mixer

As reflected from its name, LGPM is configured into a local-global modeling structure, which consists of two key branches: the spatial one for local perception and the frequency one for global perception. As shown in Fig.2, LGPM first applies a PW convolution $\widetilde{f}_{pw}$ to double the channel dimension of the input features $F_0 \in \mathbb{R}^{H\times W\times C}$. Next, the increased features $F_1 \in \mathbb{R}^{H\times W\times 2C}$ are splitted into two parts: $F_{sp} \in \mathbb{R}^{H\times W\times C}$ and $F_{fr} \in \mathbb{R}^{H\times W\times C}$, which subsequently pass through the spatial-domain branch and the frequency-domain branch, respectively. Then, we adopt the channel-attention operation to maintain the channel-level feature aggregation. Finally, another PW Conv is applied to obtain the reduced feature $F_2 \in \mathbb{R}^{H\times W\times C}$. The implementation details of spatial-domain branch and frequency-domain branch are described in the following paragraphs.

**Spatial-domain Branch:** The spatial-domain branch focuses on capturing spatial features for local and regional correlations at pixel-level. Specifically, as shown in Fig.2a, the input features $F_{sp}$ are first divided into two data-flows: $F^1_{sp}$ and $F^2_{sp}$, in which $F^1_{sp}$ remains unchanged to extract local features through a DW convolution $\widetilde{f}_{dw}$ and the other $F^2_{sp}$ is performed with pixel partition for semi-global receptive field to extract regional features. In practice, we adopt a dilated convolution $\widetilde{f}_{dc}$ to achieve regional receptive field for efficiency. Finally, we aggregate the features from the two data-flows yielding the deep spatial feature $F^d_{sp} \in \mathbb{R}^{H\times W\times C}$. The above description can be expressed as:

$$F^d_{sp} = Concat\left[\widetilde{f}_{dw}(F^1_{sp}), \widetilde{f}_{dc}(F^2_{sp})\right] \tag{1}$$

**Frequency-domain Branch:** The frequency-domain branch focuses on capturing frequency features for global modeling around the entire image. Here, we review the Fourier transform, which is widely used for analyzing the frequency characteristic of an image. Given an image $y \in \mathbb{R}^{H\times W\times C}$, the fast Fourier transform(FFT) $\mathscr{F}$ converts it to frequency space as the complex component $\mathscr{F}(y)$, which is expressed as:

$$\mathscr{F}(y)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} y(h,w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)} \tag{2}$$

where $u$ and $v$ are the coordinates of the Fourier space. $\mathscr{F}^{-1}(y)$ represents the inverse fast Fourier transform(IFFT). As demonstrated by the FFT formula Eq.2, individual components in the frequency domain correspond to specific sets of pixels in the spatial domain, necessitating distinct treatment for each frequency component. Hence, to efficiently extract unique frequency features across various degradation processes, we introduce two core components. Firstly, we implement frequency-conditional positional encoding (FCPE), which assigns a distinct identity to each frequency component. Secondly, we introduce frequency dynamic convolution (FDC), allowing flexible modeling based on the input of each frequency component. These designs utilize tailored processing methodologies to effectively leverage the unique characteristics of each frequency component.

Specifically, as shown in Fig.2b, the input features $F_{fr}$ first pass through a 2D FFT $\mathscr{F}$ to obtain the complex features: the real part $F_R \in \mathbb{R}^{H\times W\times C}$ and the imaginary part $F_I \in \mathbb{R}^{H\times W\times C}$ in frequency domain. Next, different from respectively extracting features from real and imaginary, we aggregate the real $F_R$ and imaginary $F_I$ in the channel dimension, yielding the joint features $F_J \in \mathbb{R}^{H\times W\times 2C}$. Then, $F_J$ progressively passes through the two key modules: FCPE and FDC. In FCPE, positional encoding, which is modeled by a DW Conv $\widetilde{f}_{dw}$, is incorporated into $F_J$ in the form of residual to obtain the features $F_J^{PE} \in \mathbb{R}^{H\times W\times 2C}$. The FCPE can be written as follows:

$$F_J^{PE} = F_J + \widetilde{f}_{dw}(F_J) \tag{3}$$

In FDC, we conduct a PW Conv $\widetilde{f}_{pw}$ and softmax operations to acquire dynamic weight $(w_1^{u,v}, ..., w_n^{u,v})$ for a range of learnable convolution parameters $(conv_1, ..., conv_n)$. Hence, the frequency dynamic convolution $\widetilde{f}_{fdc}^{u,v}$, at coordinates of $u, v$ in Fourier space, can be written as follows:

$$\widetilde{f}_{fdc}^{u,v} = \sum_{i=1}^{n} w_i^{u,v} \cdot conv_i \tag{4}$$

Finally, we apply the 2D IFFT to transform the modulated frequency features back to the spatial domain.

### 3.3 Multi-kernel ConvFFN

Feed-Forward Network (FN) is an effective module widely adopted in Transformers, but previous studies [83, 98] have shown that the standard FN suffers from limited capability to exploit local connection for image restoration tasks. To address this limitation, we introduce the multi-scale representation learning [76], which splits the high-dimension features into multiple pieces to extract local relationships from different receptive fields. As shown in Fig.2, MCFN first leverages a PW convolution $\widetilde{f}_{pw}$ to double the dimension of input features $F_1 \in \mathbb{R}^{H\times W\times C}$. Then, we introduce the Multi-kernel Convolution $\widetilde{f}_{MC}$ to spilt the doubled feature $F_2 \in \mathbb{R}^{H\times W\times 2C}$ into multiple heads and, in each head, we apply various kernel size DW convolutions to extract local information. Finally, after concatenating all the multi-kernel features, another PW convolution $\widetilde{f}_{pw}$ is applied to obtain the reduced feature $F_3 \in \mathbb{R}^{H\times W\times C}$. The whole operations can be expressed as:

$$F_3 = \widetilde{f}_{pw}\left(\sigma \cdot \left(\widetilde{f}_{MC}\left(\widetilde{f}_{pw}(F_1)\right)\right)\right) \tag{5}$$

Where $\sigma$ is the GELU activation function, which used after the Multi-kernel Convolution $\widetilde{f}_{MC}$ to introduce the non-linearity.

### 3.4 Loss Function

Inspired by [20, 22], we introduce a dual-domain loss in the optimization flow in keeping with the frequency-domain branch as expressed in Eq.6, which consists of two parts: a spatial domain loss and a frequency domain loss.

$$L = \left\| \hat{I} - G \right\|_1 + \lambda \left\| \mathscr{F}(\hat{I}) - \mathscr{F}(G) \right\|_1 \tag{6}$$

where L1 loss is adopted to constrain the predicted image $\hat{I}$ similar to the ground truth image $G$ and $\lambda$ is set to 0.1 for balancing dual-domain learning.

## 4 Experiments and Analysis

To evaluate our SFHformer effectiveness, we conduct extensive experiments on common image restoration tasks, including image deraining [25, 91], image dehazing [6, 38], image desnowing [15], image deblurring [2, 55], image denoising [1], image super-resolution [3], underwater image enhancement [40] and low-light image enhancement [84, 93]. In tables, the best and second-best quality scores of the evaluated methods are **highlighted** and underlined.

### 4.1 Experimental Settings

**Datasets.** We adopt RESIDE [38], O-HAZE [6], NH-HAZE [5] and DENSE-HAZE [4] for dehazing; Rain200H [91], Rain200L [91], DDN-Data [25], DID-Data [102] and SPA-Data [82] for deraining; Raindrop [63] for raindrop removal; CSD [15], SRRS [14] and Snow100K [48] for desnowing; UIEB [40] and LSUI [61] for underwater enhancement; LOL-v1 [84], LOL-v2 [93] and FiveK [9] for low-light enhancement; GoPro [55], HIDE [73] and RealBlur [72] for motion deblurring; DPDD [2] for single-image defocus deblurring; SIDD [1] for image denoising; DIV2k [3], Set5 [8], Set14 [101], B100 [52], Urban100 [30] and Manga109 [53] for efficient image super-resolution.

**Implementation Details.** AdamW optimizer [51] with $\beta_1$ and $\beta_2$ equal to 0.9 and 0.999 is used to train SFHformer. The initial learning rate is set as $10^{-3}$. We adopt the cosine annealing strategy [50] to train the models, where the learning rate gradually decreases from the initial learning rate to $10^{-6}$. All experiments are implemented by PyTorch [59] 1.7.1 with four NVIDIA 3090 GPUs. The FLOPs for all models are calculated on the input resolution of $256\times256$ for fair comparison. More implementation details for each restoration tasks can be found in Supplementary material.

### 4.2 Experimental Results

Due to limited length space, we select representative results for presentation and more comprehensive results (e.g. single-image defocus deblurring, image denoising and efficient image super-resolution) and visualizations can be detailedly found in supplementary material.

**Table 1:** Quantitative evaluations on the synthetic and real-world dehazing.

| Method | ITS [38] SOTS-indoor | | OTS [38] SOTS-outdoor | | O-HAZE [6] | | NH-HAZE [5] | | DENSE-HAZE [4] | | Overhead | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | #Param. | FLOPs |
| (TIP'16)DehazeNet [10] | 19.82 | 0.8210 | 24.75 | 0.9271 | 17.57 | 0.77 | 16.62 | 0.52 | 13.84 | 0.43 | 0.009M | 0.581G |
| (ICCV'17)AOD-Net [37] | 20.51 | 0.8164 | 24.14 | 0.9203 | 15.03 | 0.54 | 15.40 | 0.57 | 13.14 | 0.41 | 0.002M | 0.115G |
| (ICCV'19)GridDehazeNet [46] | 32.16 | 0.9845 | 30.86 | 0.9827 | 22.11 | 0.71 | 13.80 | 0.54 | - | - | 0.956M | 21.49G |
| (CVPR'20)MSBDN [23] | 33.67 | 0.9856 | 33.48 | 0.9824 | 24.36 | 0.75 | 19.23 | 0.71 | 15.37 | 0.49 | 31.35M | 41.54G |
| (AAAI'20)FFA-Net [64] | 36.39 | 0.9894 | 33.57 | 0.9842 | 22.12 | 0.77 | 19.87 | 0.69 | 14.39 | 0.45 | 4.456M | 287.8G |
| (CVPR'22)DeHamer [27] | 36.63 | 0.9881 | 35.18 | 0.9860 | 24.64 | 0.77 | 20.66 | 0.68 | 16.62 | 0.56 | 132.45M | 48.93G |
| (ECCV'22)PMNet [94] | 38.41 | 0.9900 | 34.74 | 0.9850 | 24.64 | 0.83 | 20.42 | 0.73 | 16.79 | 0.51 | 18.90M | 81.13G |
| (TIP'23)Dehazeformer [74] | 38.46 | 0.9940 | 34.29 | 0.9830 | 25.13 | 0.77 | 19.11 | 0.66 | - | - | 4.634M | 48.64G |
| (ICCV'23)FocalNet [20] | 40.82 | 0.9960 | 37.71 | 0.9950 | 25.50 | **0.94** | 20.43 | 0.79 | 17.07 | 0.63 | 3.74M | 30.63G |
| (CVPR'23)$C^2$PNet [106] | 42.56 | 0.9954 | 36.68 | 0.9900 | - | - | - | - | 16.88 | 0.57 | 7.17M | 460.9G |
| (ICCV'23)MB-TaylorFormer [65] | 42.64 | 0.9940 | 38.09 | 0.9910 | 25.31 | 0.78 | - | - | 16.44 | 0.57 | 7.43M | 88.1G |
| (Ours)SFHformer | **43.03** | **0.9966** | **38.83** | **0.9951** | **25.81** | **0.94** | **20.73** | **0.80** | **17.84** | **0.68** | 3.87M | 26.59G |

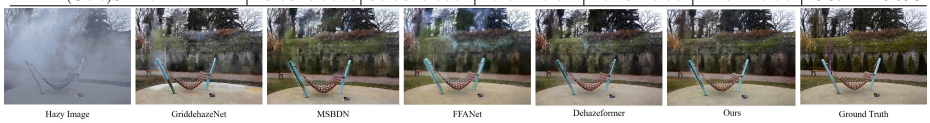


**Fig. 3:** Qualitative Results on image dehazing. Please zoom in for better comparison.

**Dehazing.** In Tab.1, Our method achieves the best performance in terms of PSNR and SSIM on all five synthetic and real-world dehazing datasets. Besides, it is worth noting that our method outperforms previous SOTA MB-TaylorFormer [65] with 0.39 dB and 0.74 dB of PSNR on the SOTS [38] indoor and outdoor sets by only 52.1% #Param. and 30.2% FLOPs. As shown in Fig.3, our method performs better color restoration and local-global level haze removal against the state-of-the-art methods.

**Deraining.** In Tab.2, Our method achieves the best performance in terms of PSNR and SSIM on all five synthetic and real-world deraining datasets. Impressively, our method outperforms previous SOTA DRSformer [16] with 0.62

**Table 2:** Quantitative evaluations on the synthetic and real-world deraining.

| Method | Rain200L [91] | | Rain200H [91] | | DID-Data [102] | | DDN-Data [25] | | SPA-Data [82] | | Overhead | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | #Param. | FLOPs |
| (ECCV'18)RESCAN [43] | 36.09 | 0.9697 | 26.75 | 0.8353 | 33.38 | 0.9417 | 31.94 | 0.9345 | 38.11 | 0.9707 | 0.150M | 32.12G |
| (CVPR'19)PReNet [69] | 37.80 | 0.9814 | 29.04 | 0.8991 | 33.17 | 0.9481 | 32.60 | 0.9459 | 40.16 | 0.9816 | 0.169M | 66.25G |
| (CVPR'20)MSPFN [33] | 38.58 | 0.9827 | 29.36 | 0.9034 | 33.72 | 0.9550 | 32.99 | 0.9333 | 43.43 | 0.9843 | 20.89M | 595.5G |
| (CVPR'20)RCDNet [80] | 39.17 | 0.9885 | 30.24 | 0.9048 | 34.08 | 0.9532 | 33.04 | 0.9472 | 43.36 | 0.9831 | 2.958M | 194.5G |
| (CVPR'21)MPRNet [100] | 39.47 | 0.9825 | 30.67 | 0.9110 | 33.99 | 0.9590 | 33.10 | 0.9347 | 43.64 | 0.9844 | 3.637M | 548.7G |
| (ICCV'21)SPDNet [96] | 40.50 | 0.9875 | 31.28 | 0.9207 | 34.57 | 0.9560 | 33.15 | 0.9457 | 43.20 | 0.9871 | 2.982M | 96.29G |
| (CVPR'22)Uformer [83] | 40.20 | 0.9860 | 30.80 | 0.9105 | 35.02 | 0.9621 | 33.95 | 0.9545 | 46.13 | 0.9913 | 20.60M | 41.09G |
| (CVPR'22)Restormer [98] | 40.99 | 0.9890 | 32.00 | 0.9329 | 35.29 | 0.9641 | 34.20 | 0.9571 | 47.98 | 0.9921 | 26.10M | 141.0G |
| (TPAMI'22)IDT [86] | 40.74 | 0.9884 | 32.10 | 0.9344 | 34.89 | 0.9623 | 33.84 | 0.9549 | 47.35 | 0.9930 | 16.39M | 58.44G |
| (CVPR'23)DRSformer [16] | 41.23 | 0.9894 | 32.18 | 0.9330 | 35.38 | 0.9647 | 34.36 | 0.9590 | 48.53 | 0.9924 | 33.70M | 242.9G |
| (Ours)SFHformer | **41.85** | **0.9908** | **32.33** | **0.9351** | **35.44** | **0.9655** | **34.38** | **0.9594** | **50.11** | **0.9942** | 7.63M | 50.59G |


**Fig. 4:** Qualitative Results on image deraining. Please zoom in for better comparison.

dB and 1.58 dB of PSNR on the rain100L [91] and SPA-Data [82] by only 22.6% #Param. and 20.8% FLOPs. As shown in Fig.4, our method restores the most similar visual quality to the ground-truth in terms of color and texture.

**Table 3:** Quantitative evaluations on motion deblurring.

| Method | GoPro [55] | | HIDE [73] | |
|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| (CVPR'20)DBGAN [104] | 31.10 | 0.942 | 28.94 | 0.915 |
| (ECCV'20)MT-RNN [58] | 31.15 | 0.945 | 29.15 | 0.918 |
| (CVPR'19)DMPHN [103] | 31.20 | 0.940 | 29.09 | 0.924 |
| (ICCV'21)SPAIR [62] | 32.06 | 0.953 | 30.29 | 0.931 |
| (ICCV'21)MIMO-UNet+ [18] | 32.45 | 0.957 | 29.99 | 0.930 |
| (CVPR'21)MPRNet [100] | 32.66 | 0.959 | 30.96 | 0.939 |
| (CVPR'22)Restormer [98] | 32.92 | 0.961 | 31.22 | 0.942 |
| (ECCV'22)Stripformer [77] | 33.08 | 0.962 | 31.03 | 0.940 |
| (ECCV'22)MPRNet-local [19] | 33.31 | 0.964 | 31.19 | 0.942 |
| (ECCV'22)Restormer-local [19] | 33.57 | 0.966 | 31.49 | 0.945 |
| (ECCV'22)NAFNet [13] | 33.71 | 0.967 | 31.31 | 0.943 |
| (ICLR'23)SFNet [22] | 33.27 | 0.963 | 31.10 | 0.941 |
| (ICML'23)IRNeXt [21] | 33.16 | 0.962 | - | - |
| (ICCV'23)icDPMs-SA [70] | 33.20 | 0.963 | 30.96 | 0.938 |
| (CVPR'23)GRL [44] | 33.93 | 0.968 | 31.62 | 0.947 |
| (Ours)SFHformer | **34.01** | **0.969** | **31.66** | **0.948** |

**Table 4:** Quantitative evaluations on low-light enhancement.

| Method | LOL-v1 [84] | | LOL-v2-real [93] | | LOL-v2-syn [93] | | Overhead | |
|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | #Param. | FLOPs |
| (CVPR'19)DeepUPE [81] | 14.38 | 0.446 | 13.27 | 0.452 | 15.08 | 0.623 | 1.02M | 21.10G |
| (AAAI'20)RF [36] | 15.23 | 0.452 | 14.05 | 0.458 | 15.97 | 0.632 | 21.54M | 46.23G |
| (CVPR'20)DeepLPF [54] | 15.28 | 0.473 | 14.10 | 0.480 | 16.02 | 0.587 | 1.77M | 5.86G |
| (CVPR'21)IPT [12] | 16.27 | 0.504 | 19.80 | 0.813 | 18.30 | 0.811 | 115.31M | 6887G |
| (CVPR'22)UFormer [83] | 16.36 | 0.771 | 18.82 | 0.771 | 19.66 | 0.871 | 20.60M | 41.09G |
| (TIP'21)Sparse [93] | 17.20 | 0.640 | 20.06 | 0.816 | 22.05 | 0.905 | 2.33M | 53.26G |
| (TIP'21)EnGAN [34] | 17.48 | 0.650 | 18.23 | 0.617 | 16.57 | 0.734 | 114.35M | 61.01G |
| (CVPR'21)RUAS [45] | 18.23 | 0.720 | 18.37 | 0.723 | 16.55 | 0.652 | 0.003M | 0.83G |
| (CVPR'20)FIDE [87] | 18.27 | 0.665 | 16.85 | 0.678 | 15.20 | 0.612 | 8.62M | 28.51G |
| (TIP'21)DRBN [92] | 20.13 | 0.830 | 20.29 | 0.831 | 23.22 | 0.927 | 5.27M | 48.61G |
| (CVPR'22)Restormer [98] | 22.43 | 0.823 | 19.94 | 0.827 | 21.41 | 0.830 | 26.10M | 141.0G |
| (ECCV'20)MIRNet [99] | 24.14 | 0.830 | 20.02 | 0.820 | 21.94 | 0.876 | 31.76M | 785G |
| (CVPR'22)SNR-Net [88] | 24.61 | 0.842 | 21.48 | 0.849 | 24.14 | 0.928 | 4.01M | 26.35G |
| (ICCV'23)Retinexformer [11] | **25.16** | 0.845 | 22.80 | 0.840 | 25.67 | 0.930 | 1.61M | 15.57G |
| (Ours)SFHformer | 24.29 | **0.862** | **23.78** | **0.872** | **25.80** | **0.937** | 1.04M | 7.75G |

**Motion Deblurring.** Tab.3 shows the quantitative results against SOTA motion deblurring methods on GoPro [55] and HIDE [73]. In practice, our model is trained only on the GoPro dataset and directly applied to the HIDE dataset. Our method achieves 1.09 dB gain in PSNR over the SOTA model Restormer [98] on GoPro [55]. Fig.5 shows visual comparisons of the evaluated models and our method generates sharper and visually-faithful results with more high frequency details. For RealBlur [72], we apply [98] settings, using a GoPro-trained model on RealBlur datasets. The detailed results are available in supplementary material.

**Low-light Enhancement.** As illustrated in Tab.4, our model surpasses most SOTA methods in terms of PSNR and SSIM on LOL-v1 [84] and LOL-v2 [93]. Particularly, our method outperforms previous SOTA Retinexformer [11] with 0.98 dB of PSNR on LOL-v2-real [93] by 64.6% #Param. and 49.8% FLOPs. As shown in Fig.6, our model achieves contrast restoration most similar to the GT, while other methods tend to exhibit either excessive brightness or darkness. Detailed results of FiveK [9] are available in supplementary material.


**Fig. 5:** Qualitative Results on image deblurring. Please zoom in for better comparison.


**Fig. 6:** Qualitative Results on low-light. Please zoom in for better comparison.

**Table 5:** Quantitative evaluations on underwater enhancement.

| Method | L-400 [61] | | | U-90 [40] | | | Overhead | |
|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | UCIQE↑ | PSNR↑ | SSIM↑ | UCIQE↑ | #Param. | FLOPs |
| (TIP'20)WaterNet [40] | 23.38 | 0.9152 | 0.5729 | 16.31 | 0.7970 | 0.5777 | 1.022M | 71.65G |
| (ICCV'21)PRWNet [31] | 27.83 | 0.9268 | 0.5801 | 20.79 | 0.8231 | 0.5830 | 2.076M | 14.71G |
| (AAAI'21)Shallow-UWNet [56] | 20.56 | 0.7675 | 0.5517 | 18.28 | 0.8553 | 0.5517 | 0.2195M | 20.14G |
| (TIP'23)U-shape Trans [61] | 24.16 | 0.9184 | 0.5871 | 21.25 | 0.8432 | 0.5882 | 22.82M | 24.35G |
| (ECCV'22)PUIE-Net [26] | 21.28 | 0.8615 | 0.5887 | 21.38 | 0.8821 | 0.5887 | 0.8322M | 21.84G |
| (TGRS'22)URSCT-SESR [71] | 29.31 | 0.9318 | 0.5890 | 22.72 | 0.9108 | **0.6140** | 11.26M | 13.84G |
| (Ours)SFHformer | **30.18** | **0.9449** | **0.5943** | **23.54** | **0.9177** | 0.6020 | 3.87M | 26.59G |

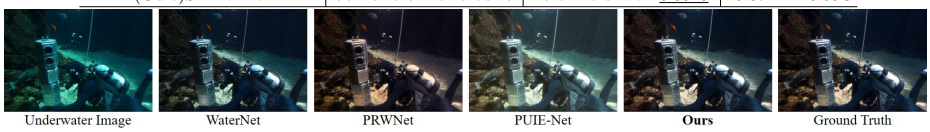


**Fig. 7:** Qualitative Results on underwater. Please zoom in for better comparison.

**Underwater Enhancement.** As illustrated in Tab.5, our model surpasses most SOTA methods in terms of PSNR, SSIM and UCIQE [90] on UIEB [40] and LSUI [61]. Particularly, our method outperforms previous SOTA URSCT-SESR [71] with 0.87 dB and 0.82 dB of PSNR on LSUI and UIEB by only 34.4% #Param. As shown in Fig.7, our method reconstructs the most similar underwater restoration images with ground-truth and achieves better recovery quality in the deepwater region.

**Table 6:** Quantitative evaluations on image desnowing.

| Method | CSD [15] | | SRRS [14] | | Snow100K [48] | | Overhead | |
|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | #Param. | FLOPs |
| (TIP'18)DesnowNet [48] | 20.13 | 0.81 | 20.38 | 0.84 | 30.50 | 0.94 | 15.60M | - |
| (CVPR'18)CycleGAN [24] | 20.98 | 0.80 | 20.21 | 0.74 | 26.81 | 0.89 | 7.840M | 42.38G |
| (CVPR'22)ALL in One [42] | 26.31 | 0.87 | 24.98 | 0.88 | 26.07 | 0.88 | 44.00M | 12.26G |
| (ECCV'20)JSTASR [14] | 27.96 | 0.88 | 25.82 | 0.89 | 23.12 | 0.86 | 65.00M | - |
| (ICCV'21)HDCW-Net [15] | 29.06 | 0.91 | 27.78 | 0.92 | 31.54 | 0.95 | 6.99M | 9.78G |
| (CVPR'22)TransWeather [79] | 31.76 | 0.93 | 28.29 | 0.92 | 31.82 | 0.93 | 21.90M | 5.64G |
| (ECCV'22)NAFNet [13] | 33.13 | 0.96 | 29.72 | 0.94 | 32.41 | **0.95** | 22.40M | 12.12G |
| (ICCV'23)FocalNet [20] | 37.18 | **0.99** | 31.34 | **0.98** | 33.53 | **0.95** | 3.74M | 30.63G |
| (ICML'23)IRNeXt [21] | 37.29 | **0.99** | 31.91 | **0.98** | 33.61 | **0.95** | 5.46M | - |
| (Ours)SFHformer | **37.45** | **0.99** | **32.39** | **0.98** | **33.94** | **0.95** | 3.87M | 26.59G |

**Table 7:** Quantitative evaluations on raindrop removal.

| Methods | Raindrop-A [63] | | Raindrop-B [63] | |
|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| (CVPR'17)pix2pix [32] | 28.02 | 0.855 | - | - |
| (CVPR'19)DuRN [47] | 31.24 | 0.926 | 25.32 | 0.817 |
| (ICCV'19)RaindropAttn [66] | 31.44 | 0.926 | - | - |
| (CVPR'18)AttentiveGAN [63] | 31.59 | 0.917 | 25.05 | 0.811 |
| (TPAMI'22)IDT [86] | 31.87 | 0.931 | - | - |
| (CVPR'22)MAXIM [78] | 31.87 | 0.935 | 25.74 | 0.827 |
| (TPAMI'23)RainDropDiff [57] | 32.43 | 0.933 | - | - |
| (Ours)SFHformer | **33.10** | **0.946** | **27.17** | **0.838** |

**Desnowing.** As shown in Tab.6, our model achieves the best performance in terms of PSNR and SSIM on three widely used desnowing datasets with less computational complexity and parameter size.

**Raindrop removal.** As shown in Tab.7, our model achieves the best performance in terms of PSNR and SSIM. Specifically, our method outperforms previous SOTA RainDropDiff [57] with 0.67 dB of PSNR on Raindrop-A.

### 4.3 Ablation Study

To present the effectiveness of our SFHformer, we conduct various ablation studies about individual components, loss function and the spatial-frequency domain hybrid structure. The evaluations are conducted on the Rain200L [91] dataset trained on image patches of 256×256.

**Table 8:** Ablation Studies on Individual Components. Baseline is in the last col.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Individual Component | FCPE | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | FDC | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | MCFN | ✓ | ✓ | | | | ✓ | ✓ | ✓ | ✓ | ✓ |
| | DFN [83] | | | ✓ | | | | | | | |
| | GDFN [98] | | | | ✓ | | | | | | |
| | MSFN [16] | | | | | ✓ | | | | | |
| | FFT Loss | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | L1 Loss | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| | Batch Norm | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ |
| | Layer Norm | | | | | | | | | ✓ | |
| Metric | PSNR↑ | 41.80 | 41.74 | 41.79 | 41.73 | 41.82 | 41.60 | 41.67 | 41.70 | 41.78 | 41.85 |
| | SSIM↑ | 0.9906 | 0.9903 | 0.9904 | 0.9903 | 0.9907 | 0.9894 | 0.9901 | 0.9903 | 0.9906 | 0.9908 |

**Individual components.** We conduct an ablation study on three proposed modules: (1) frequency-conditional positional encoding (FCPE), (2) frequency dynamic convolution (FDC), and (3) Multi-kernel ConvFFN (MCFN). As depicted in the first two rows of Table 8, removing FCPE leads to a PSNR reduction of 0.05 dB compared with the overall model, while replacing FDC with static PW Conv results in a PSNR reduction of 0.11 dB. Additionally, the third row of Table 8 illustrates that MCFN achieved PSNR gains of 0.06 dB, 0.12 dB, and 0.03 dB for DFN [83], GDFN [98], and MSFN [16], respectively.

**Effects of the dual-domain loss function.** We adopt the dual-domain loss to ensure consistency with the designed model. As shown in the fourth row of Tab.8, the dual loss obtains superior performance over single L1/FFT loss.

**Batch norm vs. Layer norm.** While maintaining the overall architecture of Transformer, we replaces the self-attention mechanism with FFT operator, rendering our model fully-convolutional. Consequently, we no longer treat images as sequential patch tokens for normalization purposes. Due to this structural shift, layer normalization, which is widely applied to sequence, is no longer suitable for our model. Instead, we opt for batch normalization. The fifth row of Tab.8 demonstrates that batch norm results in a 0.07 dB gain compared to layer norm.

**Table 9:** Ablation Studies on Spatial-domain Branch.

| structure choice | | original | dual-branch with only DW Conv | single-branch with DW Conv | single-branch with dilated conv |
|---|---|---|---|---|---|
| Metric | PSNR↑ | 41.85 | 41.81 | 41.78 | 41.76 |
| | SSIM↑ | 0.9908 | 0.9906 | 0.9906 | 0.9905 |

**Effects of the spatial-domain branch.** As shown in Tab.9, we conduct ablation study about spatial-domain branch under four different configurations. From these experiments, we derive two conclusions: firstly, the dual-branch structure outperforms the single-branch configuration; secondly, the combination of dilated convolution and depth-wise convolution yields superior results compared to using either single depth-wise convolution or single dilated convolution alone.

**Table 10:** Ablation Studies on hybrid structure.

| structure choice | | spatial-frequency hybrid | only spatial | only frequency |
|---|---|---|---|---|
| Metric | PSNR↑ | 41.85 | 41.62 | 41.71 |
| | SSIM↑ | 0.9908 | 0.9901 | 0.9902 |

**Effects of the spatial-frequency domain hybrid structure.** As shown in Tab.10, we set up 3 different structures: (1)hybrid, (2)single spatial and (3)single frequency. Tab.10 indicates that hybrid one achieves the best performance.

### 4.4 Qualitative Analyses of Frequency-domain Branch

We provide qualitative analyses of frequency-domain branch based on Fast Fourier Transform. We select four representative restoration tasks with input images sampled from GoPro [55](deblurring), LOL-v2 [93](low-light), rain200H [91](deraining) and ITS [38](dehazing). The features are obtained from the stage-5 SFHformer block in the last Local-Global Perception Mixer of decoder.


**Fig. 8:** The learned features and its FFT version generated by our frequency-domain branch for four image restoration tasks. The results are sampled from the last decoder. **(a)**degraded image; **(b)**learned features; **(c)**FFT results of (b); **(d)**ground-truth.

To substantiate the effectiveness of the proposed frequency-domain branch in modeling degradation patterns for various restoration tasks, we visualize the learned features alongside their corresponding frequency representations in Fig.8. As expected, our model adeptly captures the residual frequency patterns of various degradation (see red circle) as depicted in Fig.1c.

## 5 Conclusion

In this paper, we first rethink various degradation processes from the frequency perspective and propose an efficient transformer-like backbone, SFHformer, for image restoration. Specifically, SFHformer consists of two modules: Local-Global Perception Mixer (LGPM) and Multi-kernel ConvFFN (MCFN). Within LGPM, we innovate with a Spatial-Frequency Domain Hybrid structure, in which the spatial-domain branch emphasizes local feature extraction, while the frequency-domain branch focuses on capturing global relationships. For MCFN, multi-scale representation learning is incorporated to aggregate local and global features through multi-kernel receptive fields. Extensive experiments demonstrate state-of-the-art performance of our SFHformer on 31 benchmark datasets for a range of 10 restoration tasks.

**Acknowledgement:** This research is supported by National Natural Science Foundation of China (Grant No.62031001, Grant No.62325101).

## References


1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1692–1700 (2018) 9
2. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 111–126. Springer (2020) 1, 9
3. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017) 9
4. Ancuti, C.O., Ancuti, C., Sbert, M., Timofte, R.: Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In: 2019 IEEE international conference on image processing (ICIP). pp. 1014–1018. IEEE (2019) 9, 10
5. Ancuti, C.O., Ancuti, C., Timofte, R.: 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 workshops. pp. 444–445 (2020) 9, 10
6. Ancuti, C.O., Ancuti, C., Timofte, R., De Vleeschouwer, C.: O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 754–762 (2018) 1, 9, 10
7. Berman, D., Treibitz, T., Avidan, S.: Air-light estimation using haze-lines. In: 2017 IEEE International Conference on Computational Photography (ICCP). pp. 1–9. IEEE (2017) 2, 4
8. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012) 9
9. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: CVPR 2011. pp. 97–104. IEEE (2011) 9, 11
10. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. IEEE transactions on image processing **25**(11), 5187–5198 (2016) 10
11. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. arXiv preprint arXiv:2303.06705 (2023) 4, 5, 11
12. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12299–12310 (2021) 11
13. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European Conference on Computer Vision. pp. 17–33. Springer (2022) 11, 12
14. Chen, W.T., Fang, H.Y., Ding, J.J., Tsai, C.C., Kuo, S.Y.: Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 754–770. Springer (2020) 9, 12
15. Chen, W.T., Fang, H.Y., Hsieh, C.L., Tsai, C.C., Chen, I., Ding, J.J., Kuo, S.Y., et al.: All snow removed: Single image desnowing algorithm using hierarchical

dual-tree complex wavelet representation and contradict channel loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4196–4205 (2021) 1, 9, 12
16. Chen, X., Li, H., Li, M., Pan, J.: Learning a sparse transformer network for effective image deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5896–5905 (2023) 5, 10, 13
17. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33**, 4479–4488 (2020) 5
18. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4641–4650 (2021) 11
19. Chu, X., Chen, L., Chen, C., Lu, X.: Improving image restoration by revisiting global information aggregation. In: European Conference on Computer Vision. pp. 53–71. Springer (2022) 11
20. Cui, Y., Ren, W., Cao, X., Knoll, A.: Focal network for image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13001–13011 (2023) 9, 10, 12
21. Cui, Y., Ren, W., Yang, S., Cao, X., Knoll, A.: Irnext: Rethinking convolutional network design for image restoration (2023) 11, 12
22. Cui, Y., Tao, Y., Bing, Z., Ren, W., Gao, X., Cao, X., Huang, K., Knoll, A.: Selective frequency network for image restoration. In: The Eleventh International Conference on Learning Representations (2022) 5, 9, 11
23. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.H.: 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 (2020) 10
24. Engin, D., Genç, A., Kemal Ekenel, H.: Cycle-dehaze: Enhanced cyclegan for single image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 825–833 (2018) 12
25. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3855–3863 (2017) 9, 10
26. Fu, Z., Wang, W., Huang, Y., Ding, X., Ma, K.K.: Uncertainty inspired underwater image enhancement. In: European Conference on Computer Vision. pp. 465–482. Springer (2022) 12
27. Guo, C.L., Yan, Q., Anwar, S., Cong, R., Ren, W., Li, C.: 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 (2022) 10
28. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010) 2, 4
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 5
30. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5197–5206 (2015) 9
31. Huo, F., Li, B., Zhu, X.: Efficient wavelet boost learning-based multi-stage progressive refinement network for underwater image enhancement. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision. pp. 1944–1952 (2021) 12
32. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 12
33. Jiang, K., Wang, Z., Yi, P., Chen, C., Huang, B., Luo, Y., Ma, J., Jiang, J.: Multi-scale progressive fusion network for single image deraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8346–8355 (2020) 10
34. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. IEEE transactions on image processing **30**, 2340–2349 (2021) 11
35. Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5886–5895 (2023) 5
36. Kosugi, S., Yamasaki, T.: Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11296–11303 (2020) 11
37. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: All-in-one dehazing network. In: Proceedings of the IEEE international conference on computer vision. pp. 4770–4778 (2017) 2, 10
38. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2018) 5, 9, 10, 14
39. Li, C., Zhou, A., Yao, A.: Omni-dimensional dynamic convolution. arXiv preprint arXiv:2209.07947 (2022) 4
40. Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE Transactions on Image Processing **29**, 4376–4389 (2019) 1, 9, 12
41. Li, F., Zhang, L., Liu, Z., Lei, J., Li, Z.: Multi-frequency representation enhancement with privilege information for video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12814–12825 (2023) 5
42. Li, R., Tan, R.T., Cheong, L.F.: All in one bad weather removal using architectural search. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3175–3185 (2020) 12
43. Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 254–269 (2018) 10
44. Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: Efficient and explicit modelling of image hierarchies for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18278–18289 (2023) 5, 11
45. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10561–10570 (2021) 11
46. Liu, X., Ma, Y., Shi, Z., Chen, J.: Griddehazenet: Attention-based multi-scale network for image dehazing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7314–7323 (2019) 5, 10

47. Liu, X., Suganuma, M., Sun, Z., Okatani, T.: Dual residual networks leveraging the potential of paired operations for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7007–7016 (2019) 12
48. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. IEEE Transactions on Image Processing **27**(6), 3064–3073 (2018) 2, 9, 12
49. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 3, 5
50. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 9
51. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 9
52. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings eighth IEEE international conference on computer vision. ICCV 2001. vol. 2, pp. 416–423. IEEE (2001) 9
53. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia tools and applications **76**, 21811–21838 (2017) 9
54. Moran, S., Marza, P., McDonagh, S., Parisot, S., Slabaugh, G.: Deeplpf: Deep local parametric filters for image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12826–12835 (2020) 11
55. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017) 1, 5, 9, 11, 14
56. Naik, A., Swarnakar, A., Mittal, K.: Shallow-uwnet: Compressed model for underwater image enhancement (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 15853–15854 (2021) 4, 12
57. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 12
58. Park, D., Kang, D.U., Kim, J., Chun, S.Y.: Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In: European Conference on Computer Vision. pp. 327–343. Springer (2020) 11
59. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019) 9
60. Patro, B.N., Namboodiri, V.P., Agneeswaran, V.S.: Spectformer: Frequency and attention is what you need in a vision transformer. arXiv preprint arXiv:2304.06446 (2023) 5
61. Peng, L., Zhu, C., Bian, L.: U-shape transformer for underwater image enhancement. IEEE Transactions on Image Processing (2023) 9, 12
62. Purohit, K., Suin, M., Rajagopalan, A., Boddeti, V.N.: Spatially-adaptive image restoration using distortion-guided networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2309–2319 (2021) 11

63. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2482–2491 (2018) 9, 12
64. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11908–11915 (2020) 5, 10
65. Qiu, Y., Zhang, K., Wang, C., Luo, W., Li, H., Jin, Z.: 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 (2023) 3, 5, 10
66. Quan, Y., Deng, S., Chen, Y., Ji, H.: Deep learning for seeing through window with raindrops. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2463–2471 (2019) 12
67. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. Advances in neural information processing systems **34**, 980–993 (2021) 5
68. Rawat, W., Wang, Z.: Deep convolutional neural networks for image classification: A comprehensive review. Neural computation **29**(9), 2352–2449 (2017) 1
69. Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D.: Progressive image deraining networks: A better and simpler baseline. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3937–3946 (2019) 10
70. Ren, M., Delbracio, M., Talebi, H., Gerig, G., Milanfar, P.: Multiscale structure guided diffusion for image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10721–10733 (2023) 11
71. Ren, T., Xu, H., Jiang, G., Yu, M., Zhang, X., Wang, B., Luo, T.: Reinforced swinconvs transformer for simultaneous underwater sensing scene image enhancement and super-resolution. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–16 (2022) 12
72. Rim, J., Lee, H., Won, J., Cho, S.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 184–201. Springer (2020) 9, 11
73. Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5572–5581 (2019) 9, 11
74. Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Transactions on Image Processing **32**, 1927–1941 (2023) 2, 10
75. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2149–2159 (2022) 5
76. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 4, 8
77. Tsai, F.J., Peng, Y.T., Lin, Y.Y., Tsai, C.C., Lin, C.W.: Stripformer: Strip transformer for fast image deblurring. In: European Conference on Computer Vision. pp. 146–162. Springer (2022) 5, 11
78. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxim: Multi-axis mlp for image processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5769–5780 (2022) 12

79. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2353–2363 (2022) 12
80. Wang, H., Xie, Q., Zhao, Q., Meng, D.: A model-driven deep neural network for single image rain removal. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3103–3112 (2020) 10
81. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6849–6857 (2019) 11
82. Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W.: Spatial attentive single-image deraining with a high quality real rain dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12270–12279 (2019) 1, 9, 10, 11
83. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17683–17693 (2022) 2, 5, 8, 10, 11, 13
84. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018) 1, 5, 9, 11
85. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10551–10560 (2021) 4
86. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.J.: Image de-raining transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 10, 12
87. Xu, K., Yang, X., Yin, B., Lau, R.W.: Learning to restore low-light images via decomposition-and-enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2281–2290 (2020) 11
88. Xu, X., Wang, R., Fu, C.W., Jia, J.: Snr-aware low-light image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17714–17724 (2022) 11
89. Yang, B., Bender, G., Le, Q.V., Ngiam, J.: Condconv: Conditionally parameterized convolutions for efficient inference. Advances in neural information processing systems **32** (2019) 4
90. Yang, M., Sowmya, A.: An underwater color image quality evaluation metric. IEEE Transactions on Image Processing **24**(12), 6062–6071 (2015) 12
91. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1357–1366 (2017) 5, 9, 10, 11, 12, 14
92. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality. IEEE Transactions on Image Processing **30**, 3461–3473 (2021) 11
93. Yang, W., Wang, W., Huang, H., Wang, S., Liu, J.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. IEEE Transactions on Image Processing **30**, 2072–2086 (2021) 9, 11, 14

94. Ye, T., Zhang, Y., Jiang, M., Chen, L., Liu, Y., Chen, S., Chen, E.: Perceiving and modeling density for image dehazing. In: European Conference on Computer Vision. pp. 130–145. Springer (2022) 10
95. Yi, Q., Li, J., Dai, Q., Fang, F., Zhang, G., Zeng, T.: Structure-preserving deraining with residue channel prior guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4238–4247 (2021) 2, 4
96. Yi, Q., Li, J., Dai, Q., Fang, F., Zhang, G., Zeng, T.: Structure-preserving deraining with residue channel prior guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4238–4247 (2021) 10
97. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10819–10829 (2022) 3
98. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022) 2, 3, 4, 5, 8, 10, 11, 13
99. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Learning enriched features for real image restoration and enhancement. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 492–511. Springer (2020) 11
100. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021) 10, 11
101. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7. pp. 711–730. Springer (2012) 9
102. Zhang, H., Patel, V.M.: Density-aware single image de-raining using a multi-stream dense network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 695–704 (2018) 9, 10
103. Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5978–5986 (2019) 11
104. Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., Li, H.: Deblurring by realistic blurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2737–2746 (2020) 11
105. Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X.: Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems **30**(11), 3212–3232 (2019) 1
106. Zheng, Y., Zhan, J., He, S., Dong, J., Du, Y.: 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 (2023) 10
107. Zhou, M., Huang, J., Guo, C.L., Li, C.: Fourmer: an efficient global modeling paradigm for image restoration. In: International Conference on Machine Learning. pp. 42589–42601. PMLR (2023) 5