The Visual Computer
https://doi.org/10.1007/s00371-023-03252-8

RESEARCH

# Frequency domain-enhanced transformer for single image deraining

**Mingwen Shao[1,2] · Zhiyuan Bao[2] · Weihan Liu[2] · Yuanjian Qiao[2] · Yecong Wan[2]**

Accepted: 24 December 2023


**Abstract**
Since Transformers show a strong capability of building long-range dependencies, the relevant methods are extensively employed for image deraining tasks. However, the intrinsic limitations of Transformers, including costly computational complexity and insufficient ability to capture high-frequency components of the image, hinder the the utilization of Transformers in high-resolution images and lead to the unsatisfactory recovery of local edges and textures. To overcome these limitations, we propose an simple but effective Frequency Domain Enhanced Transformer (FDEFormer) for the image deraining. Firstly, drawing inspiration from the convolution theorem, we devise an efficient approach called frequency domain enhanced multi-head self-attention. The proposed approach replaces traditional matrix multiplication with element-wise product operations in the frequency domain, leading to a substantial reduction in computational complexity. Secondly, the existing spatial domain Transformers-based methods only focus on low-frequency features but pay less attention to high-frequency components, which can adversely affect the quality of the reconstructed images. Therefore, to integrate the content of different frequency levels, we propose a dual domain-complemented feed-forward network. Besides, we further present an attention feature fusion module to facilitate a more effective fusion of features across different layers. Extensive experiments on several datasets demonstrate that our FDEFormer performs favorably against state-of-the-art methods while taking acceptable computational costs. The source code and pre-trained models are available at https://github.com/bobozy1999/FDEFormer.

**Keywords** Image deraining · Frequency domain · Self-attention · Dual domain-complemented

## 1 Introduction

Rain is a prevalent weather occurrence that substantially impacts both the visual quality of captured rainy images and the performance of subsequent computer vision systems, including autonomous driving [1] and object detection [2]. Consequently, the removal of undesirable rain artifacts, such as rain streaks and raindrops, from a rainy image is crucial for various computer vision tasks and has garnered significant research interest in recent years.

Traditional model-based approaches [5–7] view this task as a signal separation process and investigate various priors regarding the attributes of rain streaks and raindrops. However, these manual priors are not robust enough to complex and varying rainy conditions, thus hindering the effectiveness of deraining.

With the rapid development of deep learning, deep convolutional neural networks (CNNs) -based methods show tremendous advantages in image deraining. Despite significant performance improvements, convolutional architectures have inherent limitations, i.e., convolutional operations are inherently local and spatial invariant, failing to model the spatial variation of image and thus unable to eliminate complex rain effects.

To address the above issue, transformers are applied to image deraining and achieve promising results due to their

✉ Mingwen Shao
smw278@126.com

Zhiyuan Bao
zhiyuanbao@s.upc.edu.cn

Weihan Liu
liudaneng110@126.com

Yuanjian Qiao
yjqiao@s.upc.edu.cn

Yecong Wan
yecongwan@gmail.com

[1] National Science Digital Industry College, Quanzhou Vocational and Technical University, Jinjiang 362000, Fujian, China

[2] College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266000, Shandong, China

Published online: 14 February 2024

high capacities of dynamic weighting and global dependency capturing. The majority of existing state-of-the-arts revolve around Transformers [4, 8, 9]. However, one of the biggest challenges of Transformers is the unacceptably high cost of global attention computation. To alleviate the above limitation, several effective attention mechanisms are proposed to balance efficiency and scope of dependence, such as local window attention [10], shift window attention [11] and channel attention [8]. Although promising results have been achieved, Transformers still face the following limitations. On the one hand, the high computational cost still limits the performance of Transformers for high-resolution images. On the other hand, existing Transformers are typically performed in the spatial domain, which tend to learn low-frequency components, but the capability to capture high-frequency components is insufficient (as shown in Fig. 2a, the signals in the center are more than in the edge).

In order to mitigate the above constraints, firstly, we incorporate the related calculations into the frequency domain. We note that the vanilla scaled dot-product attention [12] is employed to estimate the relevance of one token in the *query* with respect to all tokens in the *key* through the matrix multiplication. The implementation can be regarded as the convolutional operations during token rearrangement. As shown in *convolution theorem* [13], the convolution of two signals in the spatial domain is equivalent to the product of their respective Fourier transforms in the frequency domain. Based on the above theorem, we develop a practical frequency domain enhanced multi-head self-attention (FEMSA) that utilizes element-wise product operations to perform attention in the frequency domain to replace matrix multiplication. As a consequence, the computational complexity has been significantly reduced from quadratic growth ($\mathcal{O}(H^2W^2)$ for $H \times W$ feature maps) to linear growth ($\mathcal{O}(HW)$).

In addition, we observe from Fig. 2b that the residuals of degraded and clean images are distributed in the high-frequency and low-frequency regions of the spectrum, which means that the differences between degraded and clean images in the frequency domain exist simultaneously in both high-frequency and low-frequency information. Consequently, it is essential to fully leverage the information of different frequency levels for Transformers-based methods. To fuse the high-frequency and low-frequency information, we propose a straightforward yet efficient dual domain-complemented FFN (DDFN). Specifically, in addition to the conventional spatial domain nonlinear transformation, we merely incorporate another stream based on the Fast Fourier Transform (FFT) in the channel direction, named frequency stream, to supplement the missing high-frequency information in the frequency domain (Fig. 4b). This block reaps the rewards of simulating the distinctions between rainy and clean images, both in their high-frequency and low-frequency

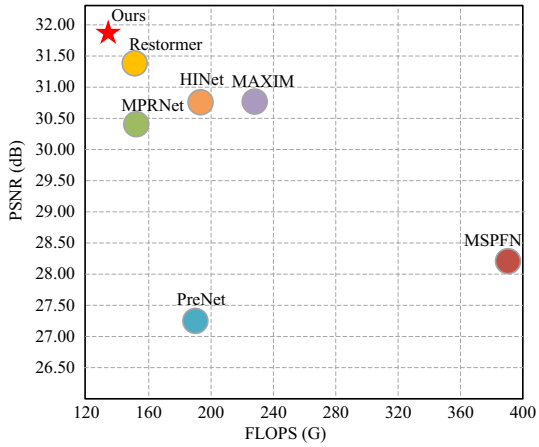



**Fig. 1** Comparisons of accuracy and floating point operations (FLOPs) with current popular deraining methods on Rain100H [3] dataset. The red star represents FDEFormer. Our FDEFormer achieves the competitive property on image deraining tasks while maintaining computationally efficient

aspects, and concurrently capturing both long-term and short-term interactions.

Also, to address the information loss in feature fusion, we further introduce a novel attention feature fusion module (AFFM, Fig. 3b) which fully utilizes the interdependence among various feature maps at different levels. Lastly, we introduce the proposed FEMSA and DDFN into the encoder and decoder architecture and formulate an asymmetric end-to-end trainable network, Frequency Domain Enhanced Transformer (FDEFormer). It should be noted that the FEMSA exclusively for the decoder module. Our experiments indicate that the proposed approach outperforms existing state-of-the-art techniques in terms of both accuracy and efficiency (Fig. 1).

The main contributions of this work are:

- We develop a highly effective frequency domain enhanced multi-head self-attention (FEMSA) which transfers the elaborate computations to the frequency domain. Our analysis demonstrates that utilizing FEMSA can significantly reduce complexity and improve performance.
- We propose a dual domain-complemented FFN (DDFN) which contains dual streams to integrate both high-frequency and low-frequency information and supply the shortcomings of the spatial domain.
- We further devise an attention feature fusion module (AFFM) for more effective feature fusion.
- Our method consistently outperforms the state-of-the-art approaches on a wide range of evaluation metrics, as demonstrated by extensive experimental results on various benchmarks.

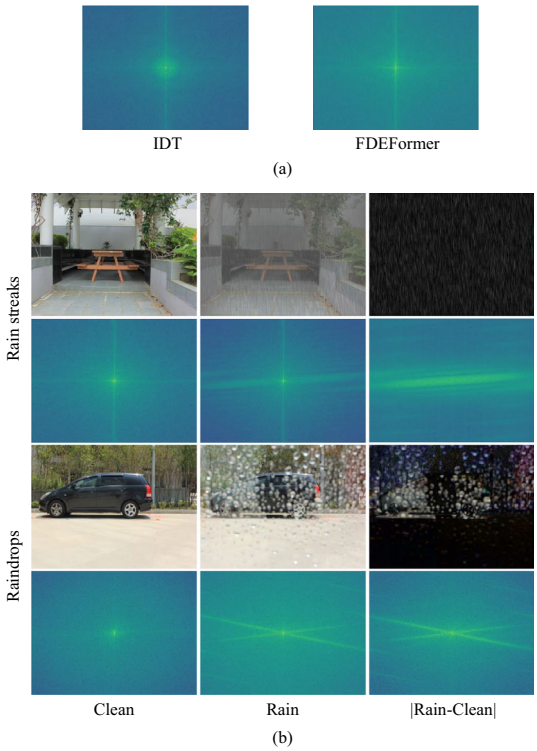



**Fig. 2** **a** Fourier spectrums of IDT [4] and FDEFormer. The Fourier spectrums are the transformation of the feature maps extracted from the same layer of the above two networks. Low-frequency signals tend to be concentrated in the center, while high-frequency signals are dispersed around the edge. **b** Visualization of Fourier spectrums between clean image and rainy image pairs. Raindrops and rain streaks are two types of degenerated phenomena. The first column shows clean images, the second column shows rainy images, and the last column demonstrates the difference in magnitude subtraction

Subsequent sections of this paper are structured in the following manners. Section 2 reviews the related work in rain removal and vision Transformers. In Sect. 3, we provide a detailed description of the implementation of our proposed method. Further, in Sect. 4, we conduct a series of comparative experiments to demonstrate the superior performance of our approach. Finally, we conclude the whole paper and give proposals for future work in Sect. 5.

## 2 Related work

In this section, we briefly introduce related advances in single image deraining and vision Transformers, respectively.

### 2.1 Single image deraining

Single image deraining refers to the process of removing undesirable rainy artifacts in the degraded image, *e.g.*, rain streaks and raindrops. The task is particularly challenging due to rain's complex and varying natures. To address this issue, there are currently two main approaches: model-based and deep learning-based methods. Model-based methods include spare coding based ones [5, 14–16], prior based ones [6, 17], and filter based ones [7, 18]. However, they rely on empirical observations and are confined to addressing specific types of rainy-day artifacts, so they may not perform as robustly in more complex and practical scenarios Deep learning-based methods [19–27] use large-scale training sets containing both rainy and clean images to learn a deraining model by exploring the properties of rain. In recent years, with the development of deep learning techniques, deep learning-based methods become the mainstream approach due to their superior generalization performance and effectiveness.

Single image deraining tasks can be further divided into two sub-tasks: rain streaks removal and raindrops removal. We briefly review related advances in the above two sub-tasks.

*Rain streaks removal* Rain streaks removal aims to remove the streaks due to the rain degradation on the camera lens or in the scene. The streaks typically exhibit directionality and may disperse attention visually, compromising image quality. During recent years, a large number of deep learning-based methods [8, 28–32] achieve great performance in removing rain streaks. To deal with dense rain accumulation and rain veiling effect in heavy rain, Li et al. [28] propose a two-stage network that integrates a physics-based backbone followed by a depth-guided GAN refinement. To enhance the utilization of information at each scale, Jiang et al. [29] propose a multi-scale progressive fusion network (MSPFN) that explores the multi-scale collaborative representation for rain streaks from the perspective of input image scales and hierarchical deep features. In order to more comprehensively investigate the features across various stages, Zamir et al. [32] propose a multi-stage architecture with a novel cross-stage feature fusion that progressively learns restoration functions for degraded inputs by optimally balancing spatial details and high-level contextualized information. Motivated by the success of fuzzy broad learning system (FBLS), Lin et al. [33] merge the merits of two-phase processing methods and FBLS and propose a novel solution for rain removal.

*Raindrops removal* Removing raindrops aims to remove the circular or irregular-shaped rain droplets on the image. This is a more challenging task since the location of raindrops and the pattern of water streaks caused by them are unpredictable and vary widely across different images. Qian et al. [34] propose an attentive generative adversarial network that injects visual attention into both the productive and discriminative networks to visually remove raindrops from a single image. In order to further utilize the deep information of the image, Peng et al. [35] propose an end-to-end mapping from a single image with raindrops to a clean image using a deep neural network with concurrent channel-spatial atten-

tion mechanism and long-short skip connections. To further improve the quality of the restored image, Shao et al. [36] fully consider the diversity of raindrops and put forward a soft mask. An iterative mechanism is introduced to extract the blur level information of raindrops, guiding their removal at different scales.

Although the mentioned deraining methods have achieved encouraging performance, they are based on CNNs and face the challenges in capturing long-range dependencies due to the limitations of convolutions. Therefore, they are unable to effectively model the global context for image deraining. To alleviate the drawback, transformers have been applied to image deraining and become the mainstream.

### 2.2 Vision transformers

Transformer [12] is a neural network structure based on self-attention mechanism, which is initially targeted at NLP task, aiming to solve the problems of long-range dependency and low computational efficiency. The remarkable achievements of Transformer garner significant interest and spurred the creation of numerous Transformer-based methods [37–39]. ViT [40] is a groundbreaking work that introduces Transformers architecture into the vision field. Converting images into multiple sequences of patches for training surpasses the state-of-the-art CNNs with remarkable results. And then, there is a surge in Transformer-based methods [39, 41, 42] applied to image restoration tasks, which achieve remarkable results. For the image super-resolution, SwinIR [11] leverages an efficient hierarchical architecture based on Swin Transformer [38]. And for the image restoration, Restormer [8] further upgrades attention block and proposes a novel Transformer architecture for multi-scale local–global representation learning on high-resolution images. Moreover, Xiao et al. [4] incorporate the notions of locality and hierarchy into the network by designing a complementary window-based Transformer and spatial Transformer to enhance locality and propose an image deraining Transformer (IDT).

The approaches showcase the potential of Transformers in tackling various image restoration challenges, and their success provides inspirations for the development of future deep learning-based models. However, the approaches are primarily conducted in the spatial domain, disregarding the frequency characteristics of images. Therefore, we consider performing calculation of Transformers in the frequency domain to reduce computational complexity while simultaneously integrating high-frequency and low-frequency information, thus compensating for the limitations of spatial domain methods.

## 3 Proposed method

Our goal is to propose an effective method for exploring the frequency domain properties of Transformers to achieve single image deraining. To this end, we first develop an efficient frequency domain enhanced multi-head self-attention (FEMSA), which calculates the attention in the frequency domain. To mitigate the limitation of Transformers in capturing high and low-frequency information, we additionally propose a dual domain-complemented feed-forward network (DDFN). In addition, we introduce an attention feature fusion module (AFFM) to adaptively fuse hierarchical features and capture learnable correlations among different layers. In this section, we first describe the overall pipeline of the proposed FDEFormer and the loss functions. Then, we present the details of each element.

### 3.1 Overall pipeline

As shown in Fig. 3, the overall structure of the proposed FDEFormer is a U-shaped hierarchical network with skip connections between the encoder and the decoder. Given the rainy image $I \in \mathbb{R}^{H\times W\times 3}$, we first apply a $3 \times 3$ convolution to extract the shallow features $X_0 \in \mathbb{R}^{H\times W\times C}$. Following the design of the U-shaped structures [8, 10], the feature maps $X_0$ pass through $K$ encoder stages. Each stage contains multiple encoder blocks and one down-sampling layer. At the beginning of each stage in the encoder (except for the first stage), we reduce the feature height and width to half while expanding the feature channel to double and then extract the deep feature through encoder blocks. Similarly, as for the decoder, we double the feature height and width while halving the feature channel at the beginning of each stage (except for the last stage). After each decoder, we fuse the features in the same stage from encoder and decoder using AFFM to acquire the deep features $X_d$ and refine them through several Transformer blocks. Ultimately, we utilize a $3\times3$ convolution to aggregate $X_d$ as a residual image, which is subsequently added to the input degraded image in order to acquire the final reconstructed image $\widehat{I}$.

Inspired by [43], We observe that the features extracted from the encoders contain more degraded information, such as raindrops and rain streaks, than those obtained from the decoder. However, the degradation can cause similar patches to change their similarity from clean features, which accordingly diminishes the result of rain removal. Therefore, we devise an asymmetric network architecture, where FEMSA is focused on the decoder stage, while the encoder stage only consists of DDFN. We refer to the Transformer blocks in the encoder stage as encoder blocks (EB) and those in the decoder stage as decoder blocks (DB).

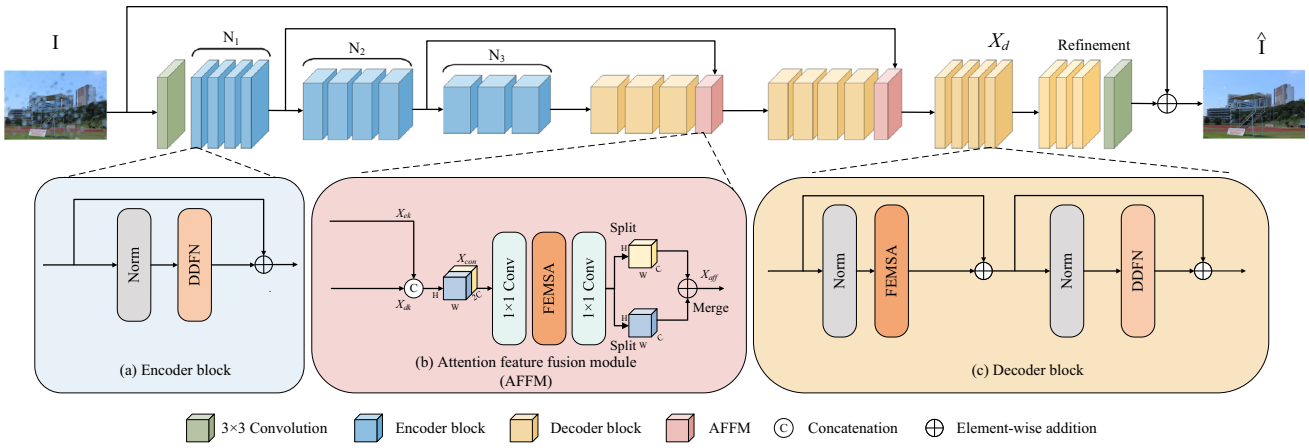



**Fig. 3** Network architectures of proposed FDEFormer. We introduce the proposed FEMSA and DDFN (details are shown in Fig. 4) into an asymmetric encoder-decoder architecture. **a** Encoder block only consists of DDFN. **b** AFFM for feature fusion. The features in the same scale from encoder and decoder are concatenated along the channel dimension and conduct feature fusion. To ensure that the channels of input and output remain equal, we perform channel-wise split. **c** Decoder block consists of both FEMSA and DDFN

We optimize FDEFormer employing the following three types of loss functions:

$$\begin{aligned} \mathcal{L}_{char} &= \sqrt{\left\|\widehat{I} - G\right\|^2 + \varepsilon^2}, \\ \mathcal{L}_{edge} &= \sqrt{\left\|\Delta\left(\widehat{I}\right) - \Delta(G)\right\|^2 + \varepsilon^2}, \\ \mathcal{L}_{fft} &= \|\mathcal{F}(\widehat{I}) - \mathcal{F}(G)\|_1, \end{aligned} \tag{1}$$

where $\widehat{I}$ is the restored image and $G$ is the corresponding ground-truth image, $\Delta$ denotes the Laplace operator [44], $\mathcal{F}$ denotes FFT operation and the constant $\varepsilon$ is empirically set to $10^{-3}$. The overall loss of the network is defined as:

$$\mathcal{L}_{All} = \mathcal{L}_{char} + \lambda_1 \mathcal{L}_{edge} + \lambda_2 \mathcal{L}_{fft}, \tag{2}$$

where $\lambda_1$ and $\lambda_2$ are hyper-parameters and set to 0.05 and 0.01 separately.

### 3.2 Frequency domain enhanced multi-head self-attention

The core design of Transformers is centered around self-attention, which also constitutes the main contributor to computational expenses. Provided with the feature $X \in \mathbb{R}^{H\times W\times C}$ with a spatial resolution of $H \times W$ pixels and $C$ channels, currently existing vision Transformers usually perform feature computation to generate *query* ($Q$), *key* ($K$) and *value* ($V$). The vanilla attention is achieved by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\alpha}\right)V, \tag{3}$$

where $Q \in \mathbb{R}^{HW\times C}$, $K \in \mathbb{R}^{HW\times C}$, $V \in \mathbb{R}^{HW\times C}$ are obtained after reshaping tensors from the original size, and $\alpha$ is a learnable scaling parameter, used to control the dot product between $Q$ and $K$. We divide the channel into multiple heads and perform parallel learning of different attention maps, following a similar approach to the conventional multi-head self-attention.

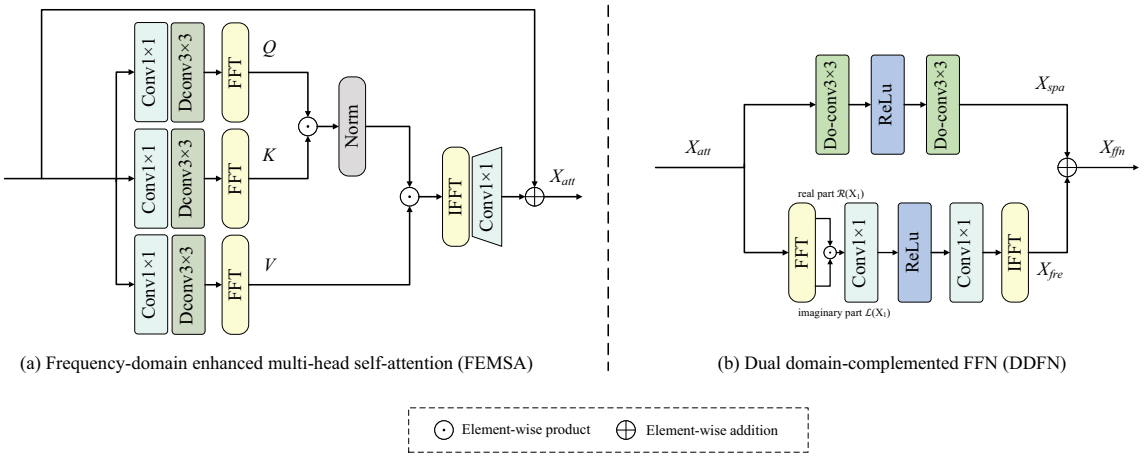
The computation of attention maps in traditional Transformer involving matrix multiplication between $Q$ and $K$ exhibits computational complexity that grows quadratically with the spatial resolution of the input, i.e., $\mathcal{O}(H^2W^2)$ for $H \times W$ feature maps. Therefore, applying self-attention to most image deraining tasks, which usually involve high-resolution images, is impractical.

From Eq. 3, taking the matrix multiplication of $Q$ and $K$ as an example, we observe that the matrix multiplication between $Q$ and $K$ is obtained via dot product:

$$\left(QK^T\right)_{ij} = Concat\langle q_i, k_j\rangle, \tag{4}$$

where $q_i$ and $k_j$ are the vectorized forms of $i^{th}$ and $j^{th}$ columns from $Q$ and $K$. We apply reshape functions to the $i^{th}$ column of $Q$ $q_i$ and all the columns $k_j$ (j = [1,...,$C$]) of $K$ (denoted as $K_{all}$), respectively. And all the $i^{th}$ column elements of $QK^T$ can be obtained by the following convolution operation as $\widehat{q_i} \otimes \widehat{K}$, where $\widehat{q_i}$ and $\widehat{K}$ denote the reshaped results of $q_i$ and $K_{all}$, and $\otimes$ denotes the convolution operation.

The *convolution theorem* [13] states that the convolution of two signals in the spatial domain is the equivalent to the product of their respective Fourier transforms in the frequency domain, which can be formulated as follows:




(a) Frequency-domain enhanced multi-head self-attention (FEMSA)

(b) Dual domain-complemented FFN (DDFN)

**Fig. 4** The sub-modules employed in the FDEFormer. **a** Frequency domain enhanced multi-head self-attention (FEMSA) that utilizes element-wise product operations to perform attention in the frequency domain. **b** Dual domain-complemented FFN (DDFN) aggregates both high-frequency and low-frequency features

$$f * g \Longleftrightarrow \mathcal{F}(f) \bullet \mathcal{F}(g), \tag{5}$$

where $*$ and $\bullet$ denote the convolution and the product operation, and $f$, $g$ are two signals.

Inspired by this, we consider replacing the matrix multiplication in the spatial domain with an element-wise product in the frequency domain, significantly reducing computational complexity. With this goal, we develop an effective frequency domain enhanced multi-head self-attention (FEMSA). To be exact, we normalize the input features and obtain $X_{norm}$. We then employ a $1 \times 1$ convolution to enhance the input feature $X_{norm} \in \mathbb{R}^{H \times W \times C}$ followed by a $3 \times 3$ depth-wise convolution to obtain features rich in local information. The output derive $Q$, $K$ and $V$, as $Q = W_d W_c X_{norm}$, $K = W_d W_c X_{norm}$ and $V = W_d W_c X_{norm}$, where $W_c$ and $W_d$ denote $1 \times 1$ convolution and $3 \times 3$ depth-wise convolution. We apply the fast Fourier transform (FFT) to the estimated features $Q$ and $K$ and calculate the element-wise product of $Q$ and $K$ in the frequency domain by:

$$A = \mathcal{F}(F_q)\overline{\mathcal{F}(F_k)}, \tag{6}$$

where $\mathcal{F}$ denotes the FFT, and $\overline{\mathcal{F}}$ denotes the conjugate transpose operation. In the same manner, the aggregation of A and V can be expressed as:

$$\widehat{A} = \mathcal{F}^{-1}\left(\text{Softmax}(A)\mathcal{F}(V)\right), \tag{7}$$

where $\mathcal{F}^{-1}$ denotes the inverse FFT. Finally, the output feature of FEMSA is generated by:

$$X_{att} = X + W_c(\widehat{A}) \tag{8}$$

where $W_c$ denotes $1 \times 1$ convolution. The detailed network architecture of the proposed FEMSA is shown in Fig. 4a.

### 3.3 Dual domain-complemented FFN

The Feed-Forward Network (FFN) can further nonlinearly transform the extracted features in the Transformer, thereby enhancing the representation and feature extraction capabilities, which is crucial for reconstructing potential clean images. We find that the Transformer model focuses more on low-frequency information while ignoring high-frequency information, resulting in semantic information loss and artifact generation. To address this issue, we introduce a novel dual domain-complemented FFN (DDFN). This block benefits from modeling the high-frequency and low-frequency differences between rainy and clean images, as well as capturing long-term and short-term interactions simultaneously. As shown in Fig. 4b, besides the normal space-domain nonlinear transformation, we only add another stream based on the FFT in the channel direction called frequency stream to supplement the missing information in the frequency domain. Let $X_{att} \in \mathbb{R}^{H \times W \times C}$ be the input feature from the attention module, where H, W, and C indicate the height, width, and channel. The FFT stream is processed as follows:

i. Apply Real FFT2d to the input tensor $X_{att}$ and obtain

$$X_i = \mathcal{F}(X_{att}) \in \mathbb{C}^{H \times W \times C},$$

ii. Concatenate the real part $\mathcal{R}(X_1)$ and the imaginary part $\mathcal{I}(X_1)$ along the channel dimension to acquire

$$X_{ii} = \mathcal{R}(X_i) \odot_C \mathcal{I}((X_i) \in \mathbb{R}^{H \times W/2 \times 2C},$$

where $\odot_C$ represents concatenation through the channel dimension;

iii. Utilize two sets of $1 \times 1$ convolution layers $\text{Conv}_1$ and $\text{Conv}_2$ with a ReLU layer in between:

$$X_{iii} = \text{Conv}_2(\text{ReLU}(\text{Conv}_1(X_{ii}))) \in \mathbb{R}^{H \times W/2 \times 2C},$$

iv. Applies inverse 2D real FFT (IFFT) to transform $X_{iii}$ back to space-domain:

$$X_{fre} = \mathcal{F}^{-1}(X_{real} + jX_{imag}) \in \mathbb{R}^{H \times W \times C},$$

where $X_{real} \in \mathbb{R}^{H \times W/2 \times C}$ and $X_{imag} \in \mathbb{R}^{H \times W/2 \times C}$ are the real part and imaginary part of $X_{iii}$, respectively, and $X_{iii} = X_{real} \odot_C X_{imag}$.

In the frequency stream, the same convolution kernels are shared across all frequencies to enable modeling the information and correlations. And the spatial stream performs nonlinear activation in the spatial domain. Ultimately, the outputs of the frequency stream are mixed with the spatial stream in together via $X_{ffn} = X_{spa} + X_{fre}$, where $X_{ffn}$, $X_{spa}$ and $X_{fre}$ denote the output of DDFN, spatial stream and frequency stream. We further substitute the convolution operation with DO-Conv [45] to attain considerable performance improvements in rain removal.

### 3.4 Attention feature fusion module

Most of the current Transformer-based methods have adopted feature concatenation or skip connections to combine features from different layers. However, the feature fusion methods can not exploit the dependency relationships among feature maps from different layers. Moreover, skip connections may suffer from information loss or vanishing gradients when the input has high resolution or depth.

In order to address the above-mentioned issues, we propose a novel attention feature fusion module (AFFM) that adaptively adjusts weights to fuse features across different layers. The AFFM architecture is shown in Fig. 3b. Given two features $X_{ek} \in \mathbb{R}^{H \times W \times C}$ and $X_{dk} \in \mathbb{R}^{H \times W \times C}$ (the $k^{th}$ encoder or decoder stage) in the same scale, we first concatenates them along the channel dimension to yield $X_{con} \in \mathbb{R}^{H \times W \times 2C}$. And then we utilize $1 \times 1$ convolutions and $3 \times 3$ depth-wise convolutions to get $Q$, $K$ and $V$. Like self-attention in FEMSA, we then reshape the $Q$, $K$ and $V$ into 2D matrices of dimensions $HW \times 2C$. We compute self-attention according to Eq. 6 and Eq. 7:

$$X_f = W_c \, \text{Attention}(Q, K, V), \tag{9}$$

where $W_c$ denotes $1 \times 1$ convolution. We finally split the output features into two equal halves along the channel

**Table 1** Comparison of average PSNR and SSIM values on RainDrop dataset [34]

| Dataset | Method / Metric | DSC [48] | pix2pix [49] | DDN [18] | AttGAN [34] | DuRN [50] | Quan et al. [51] | UMAN [36] | MSPFN [29] | MAXIM [52] | SPAIR [53] | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Test_a$ | PSNR↑ | 24.13 | 26.79 | 29.12 | 31.51 | 31.24 | 31.44 | 31.47 | 31.52 | 31.87 | 32.73 | **32.85** |
| | SSIM↑ | 0.854 | 0.864 | 0.891 | 0.901 | 0.925 | 0.926 | 0.923 | 0.929 | 0.935 | 0.941 | **0.945** |
| $Test_b$ | SSIM↑ | 23.13 | 23.50 | 24.52 | 24.92 | 25.32 | – | 25.35 | 26.56 | 25.74 | – | **27.17** |
| | SSIM↑ | 0.732 | 0.715 | 0.770 | 0.809 | 0.817 | – | 0.819 | 0.847 | 0.827 | – | **0.836** |

Best scores are **bolded** (↑: higher is better, ↓: lower is better, "–": results are not released, and the same goes for the following.). Our proposed method consistently achieves the best performance on two test datasets

**Algorithm 1** The training steps of our proposed method.

**Input:**
$I$: input rainy image;
$G$: ground-truth image;
$K$: number of encoder-decoder stages;
$k$: the $k^{th}$ encoder-decoder stage;
$\theta$: the initialization parameters of the network;
**Output:**
$\widehat{I}$: output restored image;
$\theta_{trained}$: trained deraining model;

Initialize the network parameters $\theta$: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\lambda_1 = 0.05$, $\lambda_2 = 0.01$, $batch\_size = 8$, $epoch\_num = 500$;
**for** $epoch = 1$ to $epoch\_num$ **do**
  Extracting the shallow features: $X_0 = \text{Conv}(I)$;
  **for** $k = 1$ to $K$ **do**
    FFN: $X_{ffn_{ek}} = \text{FFN}(\text{Norm}(X_{k-1}))$;
  **end for**
  **for** $k = K$ to 1 **do**
    Attention: $X_{att} = \text{FEMSA}(\text{Norm}(X_k))$;
    FFN: $X_{ffn_{dk}} = \text{FFN}(\text{Norm}(X_{att}))$;
    **if** $k = 2, 3$ **then**
      Feature fusion:
        $X_{aff} = \text{AFFM}(X_{ek}, X_{dk})$
    **end if**
  **end for**
  Refinement: $X_{re} = \text{FEMSA}(\text{Norm}(\text{FFN}(\text{Norm}(X_{FFN_{d1}}))))$
  Output: $\widehat{I} = I + \text{Conv}(X_{re})$
  Overall loss $\mathcal{L}_{All} = \mathcal{L}_{char} + \lambda_1 \mathcal{L}_{edge} + \lambda_2 \mathcal{L}_{FFT}$
  Optimize the network by minimizing $\mathcal{L}_{All}$
**end for**
**return** $\theta_{trained}$.

dimension before adding them together element-wise to obtain the fused output $X_{aff} \in \mathbb{R}^{H \times W \times C}$.

## 4 Experiments

To validate the effectiveness of our proposed method, we conduct extensive experiments on multiple datasets encompassing diverse rainy effects. Our experiments aim to demonstrate the effectiveness of the proposed method in mitigating rain-induced degradation. Additionally, we conduct detailed ablation studies and provide visualizations to offer deeper insights into the workings of the method.

### 4.1 Implementation details

FDEFormer is implemented using PyTorch and trained on a single NVIDIA RTX 3090 GPU. The specific algorithm flow is shown in Algorithm 1. The initial learning rate is set to $1 \times 10^{-4}$ and gradually decreased to $1 \times 10^{-6}$ by applying a cosine annealing strategy [46] after 200 epochs. The numbers of Transformer blocks in the FDEFormer from stage 1 to stage 3 ($N_1$ to $N_3$) are set to $\{4, 4, 10\}$, and the number of attention heads in FEMSA are $\{1, 2, 4\}$. The adam optimizer [47] is utilized with keeping $\beta_1$ and $\beta_2$ set to 0.9 and 0.999, respectively. We train our model for around 500 epochs with a batch size of 8. The input image size is randomly cropped to $256 \times 256$. For data augmentation, we perform both horizontal and vertical flipping.

### 4.2 Datasets and evaluation metrics

*Synthetic datasets* In the study of raindrops removal, we train our model on the dataset from Qian et al. (RainDrop dataset) [34], which comprises 861 paired images of clean and rainy scenes containing raindrops. The test dataset consists of two subsets, namely $Test_a$ and $Test_b$, comprising 58 and 249 samples, respectively.

For the rain streaks removal task, we train our model on the Rain13K [29, 32] dataset, which consists of a vast collection of paired images of clean and rainy scenes collected from various datasets. To evaluate the performance of our model, we utilize several synthetic datasets, including Rain100H [3], Rain100L [3], Test100 [54], Test1200 [55], and Test2800 [18].

*Real-world datasets* To validate the generalization of our method in realistic scenes, we conduct experiments on the testing publicly available real-world rainy dataset (RainDS) [56]. RainDS consists of two subsets: RainDS-Syn and RainDS-Real. RainDS-Real consists of 150 real-world images degraded by rain streaks (RS) or raindrops (RD),

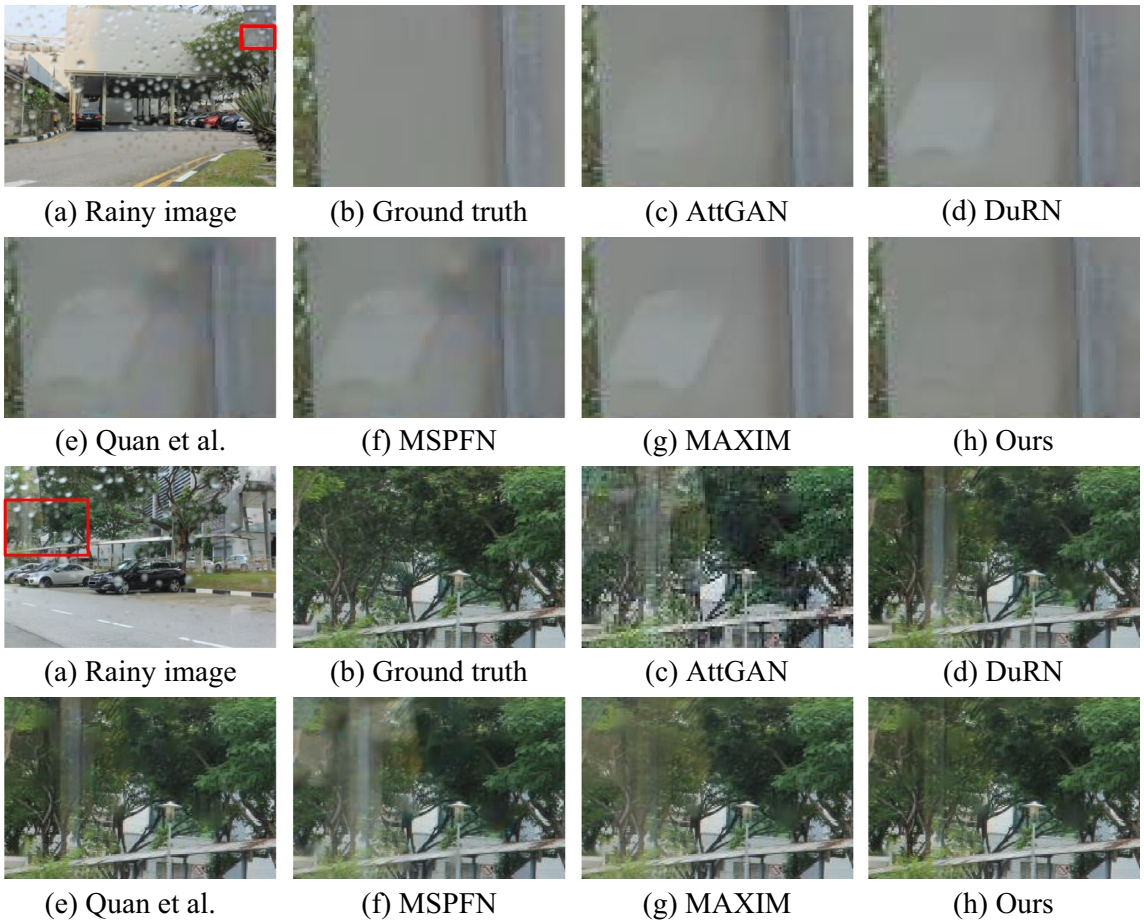



**Fig. 5** Qualitative comparisons on RainDrop dataset [34]. Our approach removes raindrops artifacts while preserving the details of the original image

along with their corresponding clean labels. All the rainy images are collected using the DSLR camera.

*Evaluation metrics* In order to evaluate the performance of the proposed model, we employ two standard image quality metrics, PSNR [57] and SSIM [58] for the above benchmarks. Following the previous methods [29, 53, 59], we compute PSNR and SSIM on Y channels of YCbCr space.

### 4.3 Comparison with the state-of-the-arts

*Quantitative comparison* The quantitative results regarding PSNR and SSIM for raindrops removal and rain streaks removal are reported in Tables 1 and Table 2, respectively. As shown in Table 1, FDEFormer outperforms the current deraining model significantly for removal of raindrops. FDEFormer surpasses the best existing model by 0.12dB and 1.43dB in terms of PSNR on $Test_a$ and $Test_b$, respectively. And for the rain streaks removal task, FDEFormer achieves competitive performances on each dataset, where state-of-the-arts are achieved on the Rain100H [3] and Test1200 [55]. Although the results on the part of the datasets are lower than Restormer [8], FDEFormer shows more wonderful performance in computing consumption and reasoning time, as shown in Fig. 1 and Table 8.

To corroborate the superiority of FDEFormer, we assess the performance on the demanding dataset RainDS [56]. As indicated in Table 3, FDEFormer achieves remarkable improvements over the state-of-the-arts methods. It can be seen that FDEFormer attains the best results on all types of rain effects for both the rain streaks subset (RS) and the raindrops subset (RD). These comprehensive results provide empirical evidence that our proposed method has the potential to achieve superior performance compared to existing methods for the task of single image deraining.

*Qualitative comparison* We present qualitative comparisons of deraining results on multiple rainy cases, including raindrops (Fig. 5), rain streaks (Fig. 6) and real-world scenes (Fig. 7), to verify the effectiveness of FDEFormer. For instance, the first row in Fig. 6 presents an example chosen from Rain100H [3]. Rainy image contains rain streaks which are similar to the fences on the ground. Other methods either fail to thoroughly remove rain streaks or inadver-

**Table 2** Quantitative comparisons with the state-of-the-arts on the Rain13K [29, 32]

| | Test100 [54] | | Rain100H [3] | | Rain100L [3] | | Test2800 [18] | | Test1200 [55] | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| DerainNet [60] | 22.77 | 0.810 | 14.92 | 0.592 | 27.03 | 0.884 | 24.31 | 0.861 | 23.38 | 0.835 | 22.48 | 0.796 |
| SEMI [61] | 22.35 | 0.788 | 16.56 | 0.486 | 25.03 | 0.842 | 24.43 | 0.782 | 26.05 | 0.822 | 22.88 | 0.744 |
| DIDMDN [55] | 22.56 | 0.818 | 17.35 | 0.524 | 25.23 | 0.741 | 28.13 | 0.867 | 29.65 | 0.901 | 24.58 | 0.770 |
| UMRL [62] | 24.41 | 0.829 | 26.01 | 0.832 | 29.18 | 0.923 | 29.97 | 0.905 | 30.55 | 0.910 | 28.02 | 0.880 |
| RESCAN [63] | 25.00 | 0.835 | 26.36 | 0.786 | 29.80 | 0.881 | 31.29 | 0.904 | 30.51 | 0.882 | 28.59 | 0.857 |
| PreNet [64] | 24.81 | 0.851 | 26.77 | 0.858 | 32.44 | 0.950 | 31.75 | 0.916 | 31.36 | 0.911 | 29.42 | 0.897 |
| MSPFN [29] | 27.50 | 0.876 | 28.66 | 0.860 | 32.40 | 0.933 | 32.82 | 0.930 | 32.39 | 0.916 | 30.75 | 0.903 |
| MPRNet [32] | 30.27 | 0.897 | 30.41 | 0.890 | 36.40 | 0.965 | 33.64 | 0.938 | 32.91 | 0.916 | 32.73 | 0.921 |
| SPAIR [53] | 30.35 | 0.909 | 30.95 | 0.892 | 36.93 | 0.969 | 33.34 | 0.936 | 33.04 | 0.922 | 32.91 | 0.926 |
| HINet [65] | 30.29 | 0.906 | 30.65 | 0.894 | 37.28 | 0.970 | 33.91 | 0.941 | 33.05 | 0.919 | 33.03 | 0.926 |
| MAXIM [52] | 31.17 | 0.922 | 30.81 | 0.903 | 38.06 | 0.977 | 33.80 | 0.943 | 32.37 | 0.922 | 33.24 | 0.933 |
| IDT* [4] | 31.30 | 0.917 | 31.02 | 0.894 | 38.32 | 0.972 | 33.92 | 0.939 | 32.58 | 0.925 | 33.43 | 0.930 |
| Restormer [8] | **32.00** | **0.923** | 31.46 | 0.904 | **38.99** | **0.978** | **34.18** | 0.944 | 33.19 | 0.926 | **33.96** | 0.935 |
| Ours | 31.54 | **0.923** | **31.87** | **0.907** | 38.65 | 0.977 | 34.14 | **0.945** | **33.33** | **0.945** | 33.91 | **0.939** |

∗ denotes that the official source code is not released, but we reproduce the method ourselves. Best scores are **bolded** and second-best are underlined. Our proposed method achieves the best or second-best results on all five test datasets

**Table 3** Quantitative results on the real-world dataset RainDS [56]

| Method | RS | | RD | |
|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| MSPFN [29] | 24.76 | 0.691 | 23.02 | 0.606 |
| MPRNet [32] | 24.77 | 0.700 | 23.20 | 0.618 |
| HINet [65] | 24.71 | 0.693 | 23.42 | 0.644 |
| Restormer [8] | 24.73 | 0.689 | 23.49 | 0.659 |
| Ours | 24.98 | 0.709 | 23.55 | 0.668 |

The proposed FDEFormer attains best results on all types of rain effects, i.e., rain streaks (RS), and raindrops (RD)

tently remove the fences as well. Due to its powerful dual domain integration feature, FDEFormer can extract more discriminative features from both local and global regions. Consequently, it attains the optimal visualization outcome of effectively removing rain streaks while preserving a greater degree of image information. To evaluate the generalization of model, we also estimate the visual effect on the real-world dataset RainDS. As illustrated in Fig. 7, when compared to other techniques, FDEFormer produces the most aesthetically pleasing outcome.

### 4.4 Ablation studies

We perform comprehensive ablations to demonstrate the individual contributions of each component in FDEFormer toward its overall effectiveness. All evaluations are performed on RainDrop dataset. We train multiple models on image patches of size 256 × 256 and test on the $Test_a$ dataset uniformly. All experiments are performed under the same conditions except for the removed modules in each experiment.

*Effect of FEMSA* The proposed FEMSA is used to reduce the computational cost. To verify whether the attention performed in the frequency domain impacts the result, we compare the FEMSA with the baseline method that acts in the spatial domain (Attention+DDFN). Table 4 shows the quantitative evaluation results. The method that computes attention in the spatial domain cannot generate good deraining results, where its PSNR value is 0.50 lower (see comparisons of "Attention+DDFN" and "FEMSA+DDFN" in Table 4). Moreover, compared to the baseline method only using FFN ("w/ only FFN"), using the proposed FEMSA in this baseline generates much better results, where the PSNR value is 0.75dB higher (see comparisons of "w/ only FFN" and "FEMSA+FFN" in Table 4.

*Effect of DDFN* To demonstrate the effectiveness of DDFN, we perform two groups of ablations. For the first one, we compare the proposed method only using the FEMSA and DDFN (FEMSA+DDFN) with the method using the FEMSA and original FFN (FEMSA+FFN). Table 4 show that using

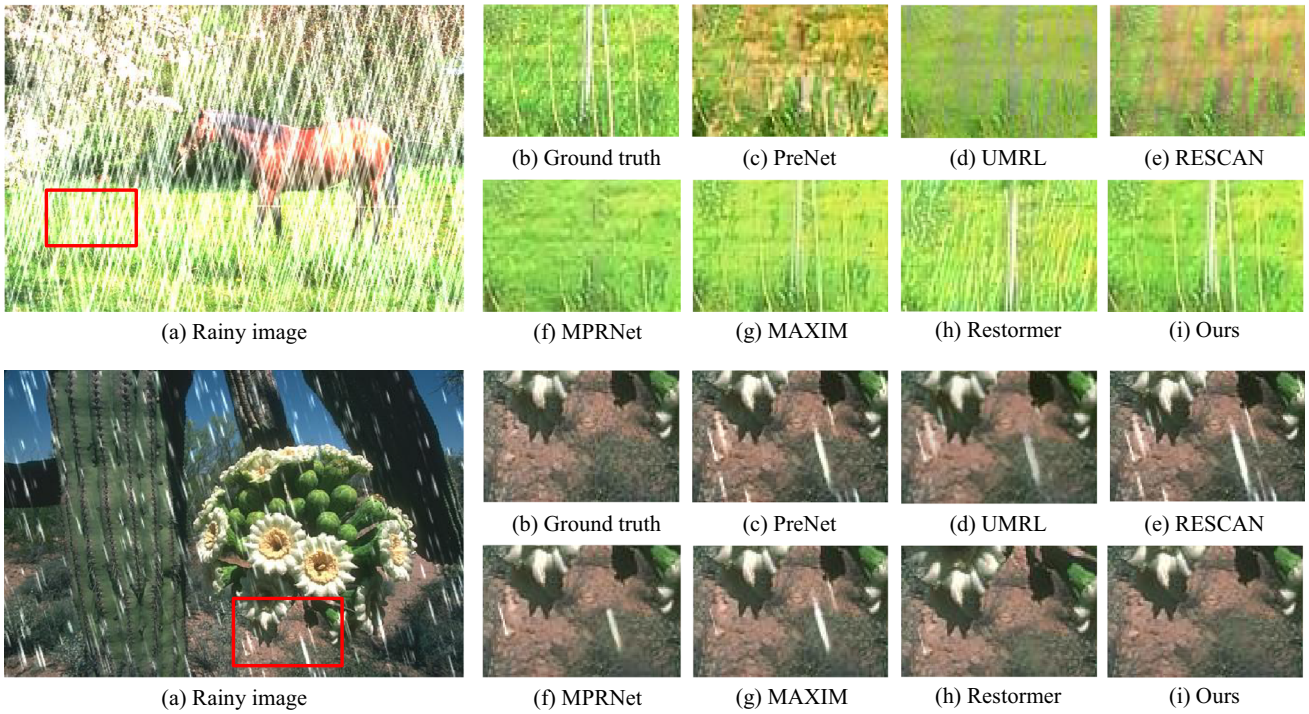



**Fig. 6** Qualitative comparisons of Rain100H and Rain100L [3]. Our approach resulted in cleaner and more visually compelling results

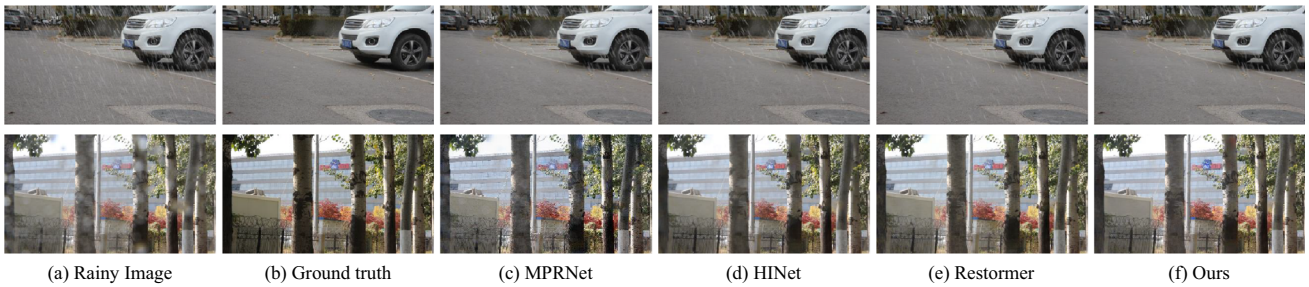



**Fig. 7** Qualitative comparisons on real-world rainy images from RainDS [56]. Compared to other methods, our approach yielded superior visual results


**Fig. 8** Visual comparison results of ablation studies with different module settings

**Table 4** The contributions of FEMSA and DDFN. "✔" indicates the presence of the module in the network and "✘" means not used

| No | Attention | FEMSA | FFN | DDFN | PSNR↑ | SSIM↑ |
|---|---|---|---|---|---|---|
| w/ only FFN | ✘ | ✘ | ✔ | ✘ | 31.95 | 0.922 |
| w/ only DDFN | ✘ | ✘ | ✘ | ✔ | 32.22 | 0.930 |
| Attention+DDFN | ✔ | ✘ | ✘ | ✔ | 32.35 | 0.934 |
| FEMSA+FFN | ✘ | ✔ | ✔ | ✘ | 32.70 | 0.939 |
| FEMSA+DDFN | ✘ | ✔ | ✘ | ✔ | **32.85** | **0.945** |

The best results are achieved when both FEMSA and DDFN are employed together
Best scores are highlighted in bold

**Table 5** Ablation analysis on AFFM. Experimental results demonstrate that AFFM outperforms other connections

| Methods | Feature concatenation | Skip connection | AFFM (Ours) |
|---|---|---|---|
| PSNR↑ | 31.78 | 32.20 | **32.85** |
| SSIM↑ | 0.918 | 0.931 | **0.945** |

Best scores are highlighted in bold

**Table 6** Quantitative evaluations of the asymmetric encoder-decoder network design

| Methods | FEMSA in enc&dec | FEMSA only in dec (Ours) |
|---|---|---|
| PSNR↑ | 32.56 | **32.85** |
| SSIM↑ | 0.938 | **0.945** |

Best scores are highlighted in bold

**Table 7** Ablation studies of the loss function. The experiment achieves the best results when all three losses are present

| Loss | PSNR↑ | SSIM↑ |
|---|---|---|
| $L_{char}$ | 32.29 | 0.926 |
| $L_{char}+\lambda_1 L_{edge}$ | 32.53 | 0.932 |
| $L_{char}+\lambda_1 L_{edge}+\lambda_2 L_{FFT}$(ours) | **32.85** | **0.945** |

Best scores are highlighted in bold

the proposed DDFN generates better results, where the PSNR value is 0.15dB higher. For the second, the comparisons of "w/ only FFN" and "w/ only DDFN" show that using the proposed DDFN further verifies the effect. The visual comparison results of the ablation of FEMSA and DDFN are shown in Fig. 8.

*Effect of AFFM* We replace the AFFM with feature concatenations and skip connections, respectively, and set two baselines. Table 5 shows that the AFFM provides better results.

*Design of asymmetric architecture* To examine the effect of this asymmetric architecture design, we compare the network that puts the FEMSA into both the encoder and decoder modules (FEMSA in enc&dec). Table 6 shows that using the FEMSA only in the decoder module generates better results, where the PSNR value is 0.39dB higher.

*Loss functions* We conduct additional ablation studies on the loss function presented in Table 7. The addition of FFT Loss results in a marked improvement in the quality of reconstruction. Our findings demonstrate that this novel loss function is capable of enhancing the authenticity of the restored image.

*Model efficiency* Figure 1 and Table 8 illustrate the contrast of the FLOPs with current popular image deraining methods. It can be seen that our method outperforms all other methods currently available with minimal calculation consumption. It proves that our proposed method is not only effective but also possesses high application values.

## 5 Conclusion

In this paper, we present a simple but effective method to explore the performance of Transformers in the frequency domain for rain removal. We introduce key design elements into the core components of Transformer blocks to reduce computation complexity and further enhance performance. Specifically, the frequency domain enhanced multi-head self-attention (FEMSA) utilizes element-wise product operations to perform attention in the frequency domain to reduce computational complexity. Furthermore, the proposed dual domain-complemented FFN (DDFN) models the high-frequency and low-frequency differences between rainy and clean images simultaneously to address the limitations of spatial domain Transformers. Moreover, we design an asymmetric network architecture and incorporate the attention feature fusion module (AFFM). Extensive quantitative and qualitative experiments on synthetic and real-world datasets demonstrate the availability and superiority of the proposed approach. However, our method has limitations: it has a relatively large number of parameters due to the characteristics of Transformers. In our future research, we will implement either pruning or distillation techniques in our model to preserve the original deraining performance while attaining trustworthy model compression.

**Table 8** Comparison of computational complexity among different models

| Methods | MSPFN [29] | MAXIM [52] | HINet [65] | PreNet [64] | MPRNet [32] | Restormer [8] | Ours |
|---|---|---|---|---|---|---|---|
| FLOPS (G)↓ | 391.93 | 216.03 | 170.71 | 168.96 | 141.28 | 140.99 | **131.53** |

**Acknowledgements** The authors are very indebted to the anonymous referees for their critical comments and suggestions for the improvement of this paper. This work was supported by National Key Research and development Program of China (2021YFA1000102), and in part by the grants from the National Natural Science Foundation of China (Nos. 62376285, 62272375, 61673396), Natural Science Foundation of Shandong Province, China (No. ZR2022MF260).

**Author Contributions** All authors did not receive support from any organization for the submitted work. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Weihan Liu, Yuanjian Qiao and Yecong Wan. The first draft of the manuscript was written by Mingwen Shao and Zhiyuan Bao and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

## Declarations

**Conflict of interest** All authors declare that they have no conflict of interest.

## References

1. Sun, H., Ang, M.H., Rus, D.: A convolutional network for joint deraining and dehazing from a single image for autonomous driving in rain. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 962–969 (2019). https://doi.org/10.1109/IROS40897.2019.8967644
2. Wang, W., Zhang, J., Zhai, W., Cao, Y., Tao, D.: Robust object detection via adversarial novel style exploration. IEEE Trans. Image Process. **31**, 1949–1962 (2022). https://doi.org/10.1109/TIP.2022.3146017
3. 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). https://doi.org/10.1109/CVPR.2017.183
4. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.-J.: Image de-raining transformer. IEEE Trans. Pattern Anal. Mach. Intell., 1–18 (2022). https://doi.org/10.1109/TPAMI.2022.3183612
5. Deng, L., Huang, T., Zhao, X., Jiang, T.: A directional global sparse model for single image rain removal. Appl. Math. Model. **59**, 662–679 (2018). https://doi.org/10.1016/j.apm.2018.03.001
6. Xu, J., Zhao, W., Liu, P., Tang, X.: Removing rain and snow in a single image using guided filter. In: 2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE), vol. 2, pp. 304–307 (2012). https://doi.org/10.1109/CSAE.2012.6272780
7. Zheng, X., Liao, Y., Guo, W., Fu, X., Ding, X.: Single-image-based rain and snow removal using multi-guided filter. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3–7, 2013. Proceedings, Part III 20, pp. 258–265 (2013). https://doi.org/10.1007/978-3-642-42051-1-33
8. 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). https://doi.org/10.1109/CVPR52688.2022.00564
9. Wan, Y., Shao, M., Bao, Z., Cheng, Y.: Global-local transformer for single-image rain removal. Pattern Anal. Appl. 1–12 (2023). https://doi.org/10.1007/s10044-023-01184-6
10. 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). https://doi.org/10.1109/CVPR52688.2022.01716
11. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833–1844 (2021). https://doi.org/10.1109/ICCVW54120.2021.00210
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
13. Blackledge, J.M.: Digital Image Processing: Mathematical and Computational Methods, pp. 44–45. Elsevier, England (2005)
14. Wang, Y., Liu, S., Chen, C., Zeng, B.: A hierarchical approach for rain or snow removing in a single color image. IEEE Trans. Image Process. **26**(8), 3936–3950 (2017). https://doi.org/10.1109/TIP.2017.2708502
15. Chen, D., Chen, C., Kang, L.: Visual depth guided color image rain streaks removal using sparse coding. IEEE Trans. Circuits Syst. Video Technol. **24**(8), 1430–1455 (2014). https://doi.org/10.1109/TCSVT.2014.2308627
16. Zhu, L., Fu, C., Lischinski, D., Heng, P.: Joint bi-layer optimization for single-image rain streak removal. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2526–2534 (2017). https://doi.org/10.1109/ICCV.2017.276
17. Jiang, T., Huang, T., Zhao, X., Deng, L., Wang, Y.: Fastderain: a novel video rain streak removal method using directional gradient priors. IEEE Trans. Image Process. **28**(4), 2089–2102 (2018). https://doi.org/10.1109/TIP.2018.2880512
18. 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). https://doi.org/10.1109/CVPR.2017.186
19. Chen, X., Pan, J., Jiang, K., Li, Y., Huang, Y., Kong, C., Dai, L., Fan, Z.: Unpaired deep image deraining using dual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2017–2026 (2022). https://doi.org/10.1109/CVPR52688.2022.00206
20. Du, Y., Xu, J., Zhen, X., Cheng, M.-M., Shao, L.: Conditional variational image deraining. IEEE Trans. Image Process. **29**, 6288–6301 (2020). https://doi.org/10.1109/TIP.2020.2990606
21. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17452–17462 (2022). https://doi.org/10.1109/CVPR52688.2022.01693
22. Wan, Y., Shao, M., Cheng, Y., Liu, Y., Bao, Z., Meng, D.: Restoring images captured in arbitrary hybrid adverse weather conditions in one go. arXiv preprint arXiv:2305.09996 (2023)
23. Wan, Y., Cheng, Y., Shao, M., Gonzàlez, J.: Image rain removal and illumination enhancement done in one go. Knowl.-Based Syst. **252**, 109244 (2022). https://doi.org/10.1016/j.knosys.2022.109244
24. Shao, M., Qiao, Y., Meng, D., Zuo, W.: Uncertainty-guided hierarchical frequency domain transformer for image restoration. Knowl.-Based Syst. **263**, 110306 (2023). https://doi.org/10.1016/j.knosys.2023.110306
25. Yang, H., Zhou, D., Li, M., Zhao, Q.: A two-stage network with wavelet transformation for single-image deraining. Vis. Comput., 1pp. –17 (2022). https://doi.org/10.1007/s00371-022-02533-yv

26. Chen, M., Wang, P., Shang, D., Wang, P.: Cycle-attention-derain: unsupervised rain removal with cyclegan. Vis. Comput., 1–13 (2023). https://doi.org/10.1007/s00371-023-02947-2
27. Luo, Y., Wu, M., Huang, Q., Zhu, J., Ling, J., Sheng, B.: Joint feedback and recurrent deraining network with ensemble learning. Vis. Comput. **38**(9–10), 3109–3119 (2022). https://doi.org/10.1007/s00371-022-02567-2
28. Li, R., Cheong, L.-F., Tan, R.T.: Heavy rain image restoration: integrating physics model and conditional adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1633–1642 (2019). https://doi.org/10.1109/CVPR.2019.00173
29. 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). https://doi.org/10.1109/CVPR42600.2020.00837
30. Huang, H., Yu, A., He, R.: Memory oriented transfer learning for semi-supervised image deraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7732–7741 (2021). https://doi.org/10.1109/CVPR46437.2021.00764
31. Yasarla, R., Sindagi, V.A., Patel, V.M.: Syn2real transfer learning for image deraining using gaussian processes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2726–2736 (2020). https://doi.org/10.1109/CVPR42600.2020.00280
32. 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). https://doi.org/10.1109/CVPR46437.2021.01458
33. Lin, X., Ma, L., Sheng, B., Wang, Z.-J., Chen, W.: Utilizing two-phase processing with fbls for single image deraining. IEEE Trans. Multimedia **23**, 664–676 (2020). https://doi.org/10.1109/TMM.2020.2987703
34. 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). https://doi.org/10.1109/CVPR.2018.00263
35. Peng, J., Xu, Y., Chen, T., Huang, Y.: Single-image raindrop removal using concurrent channel-spatial attention and long-short skip connections. Pattern Recognit. Lett. **131**, 121–127 (2020). https://doi.org/10.1016/j.patrec.2019.12.012
36. Shao, M., Li, L., Meng, D., Zuo, W.: Uncertainty guided multi-scale attention network for raindrop removal from a single image. IEEE Trans. Image Process. **30**, 4828–4839 (2021). https://doi.org/10.1109/TIP.2021.3076283
37. Lin, X., Sun, S., Huang, W., Sheng, B., Li, P., Feng, D.D.: Eapt: efficient attention pyramid transformer for image processing. IEEE Trans. Multimedia (2021)
38. 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). https://doi.org/10.1109/ICCV48922.2021.00986
39. 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). https://doi.org/10.1109/CVPR46437.2021.01212
40. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
41. 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). https://doi.org/10.1109/CVPR52729.2023.01753
42. Chen, H., Gu, J., Liu, Y., Magid, S.A., Dong, C., Wang, Q., Pfister, H., Zhu, L.: Masked image training for generalizable deep image denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1692–1703 (2023). https://doi.org/10.1109/CVPR52729.2023.00169
43. 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). https://doi.org/10.1109/CVPR52729.2023.00570
44. Kamgar-Parsi, B., Rosenfeld, A.: Optimally isotropic Laplacian operator. IEEE Trans. Image Process. **8**(10), 1467–1472 (1999). https://doi.org/10.1109/83.791975
45. Cao, J., Li, Y., Sun, M., Chen, Y., Lischinski, D., Cohen-Or, D., Chen, B., Tu, C.: Do-conv: depthwise over-parameterized convolutional layer. IEEE Trans. Image Process. **31**, 3726–3736 (2022). https://doi.org/10.1109/TIP.2022.3175432
46. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
47. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 **3** (2014)
48. Luo, Y., Xu, Y., Ji, H.: Removing rain from a single image via discriminative sparse coding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3397–3405 (2015). https://doi.org/10.1109/ICCV.2015.388
49. 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). https://doi.org/10.1109/CVPR.2017.632
50. 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). https://doi.org/10.1109/CVPR.2019.00717
51. 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). https://doi.org/10.1109/ICCV.2019.00255
52. 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). https://doi.org/10.1109/CVPR52688.2022.00568
53. 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). https://doi.org/10.1109/ICCV48922.2021.00231
54. Zhang, H., Sindagi, V., Patel, V.M.: Image de-raining using a conditional generative adversarial network. IEEE Trans. Circuits Syst. Video Technol. **30**(11), 3943–3956 (2019). https://doi.org/10.1109/TCSVT.2019.2920407
55. 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). https://doi.org/10.1109/CVPR.2018.00079
56. Quan, R., Yu, X., Liang, Y., Yang, Y.: Removing raindrops and rain streaks in one go. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9147–9156 (2021). https://doi.org/10.1109/CVPR46437.2021.00903

57. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electron. Lett. **44**(13), 800–801 (2008). https://doi.org/10.1049/el:20080522
58. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
59. Fu, X., Xiao, J., Zhu, Y., Liu, A., Wu, F., Zha, Z.-J.: Continual image deraining with hypergraph convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. (2023). https://doi.org/10.1109/TPAMI.2023.3241756
60. Fu, X., Huang, J., Ding, X., Liao, Y., Paisley, J.: Clearing the skies: a deep network architecture for single-image rain removal. IEEE Trans. Image Process. **26**(6), 2944–2956 (2017). https://doi.org/10.1109/TIP.2017.2691802
61. Wei, W., Meng, D., Zhao, Q., Xu, Z., Wu, Y.: Semi-supervised transfer learning for image rain removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3877–3886 (2019). https://doi.org/10.1109/CVPR.2019.00400
62. Yasarla, R., Patel, V.M.: Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8405–8414 (2019). https://doi.org/10.1109/CVPR.2019.00860
63. 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). https://doi.org/10.1007/978-3-030-01234-2-16
64. 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). https://doi.org/10.1109/CVPR.2019.00406
65. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 182–192 (2021). https://doi.org/10.1109/CVPRW53098.2021.00027




**Mingwen Shao** received his M.S. degree in mathematics from the Guangxi University, Guangxi, China, in 2002, and the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2005. He received the postdoctoral degree in control science and engineering from Tsinghua University in February 2008. Now he is a professor and doctoral supervisor at China University of Petroleum (East China). His research interests include machine learning, computer vision, and data mining.

**Zhiyuan Bao** is a master student at College of Computer Science and Technology, China University of Petroleum (East China), under the supervision of Prof. Shao. He received the bachelor of software engineering from the China University of Petroleum (East China). His research interests include image restoration and computer vision.

**Weihan Liu** is a master's student at the College of Computer Science and Technology, China University of Petroleum (East China), under the supervision of Professor Shao. He holds a bachelor's degree in computer science and technology from the Qilu University of Technology (Shandong Academy of Sciences). His research focuses on image restoration and computer vision.

**Yuanjian Qiao** received the M.S. degree in electrical engineering and automation from the Qilu University of Technology (Shandong Academy of Sciences), Jinnan, China, in 2021. He is currently pursuing the Ph.D. degree under the supervision of Prof. M. Shao in the School of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China. His current research interests include image restoration and deep learning.

**Yecong Wan** received the B.Eng. degrees in College of Computer Science and Technology, China University of Petroleum, City Qingdao, China. Now he is an M.S. at China University of Petroleum (East China), under the supervision of Prof. Shao. His current research interests include image restoration and computer vision.