Efficient Frequency-Domain Image Deraining with Contrastive Regularization

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

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

Abstract. Most current single image-deraining (SID) methods are based on the Transformer with global modeling for high-quality reconstruction. However, their architectures only build long-range features from the spatial domain, which suffers from a significant computational burden to keep effectiveness. Besides, these methods either overlook negative sample information in training or underutilize the rain streak patterns present in the negative ones. To tackle these problems, we propose a Frequency-Aware Deraining Transformer Framework (FADformer) that fully captures frequency domain features for efficient rain removal. Specifically, we construct the FADBlock, including the Fused Fourier Convolution Mixer (FFCM) and Prior-Gated Feed-forward Network (PGFN). Unlike self-attention mechanisms, the FFCM conducts convolution operations in both spatial and frequency domains, endowing it with local-global capturing capabilities and efficiency. Simultaneously, the PGFN introduces residue channel prior in a gating manner to enhance local details and retain feature structure. Furthermore, we introduce a Frequency-domain Contrastive Regularization (FCR) during training. The FCR facilitates contrastive learning in the frequency domain and leverages rain streak patterns in negative samples to improve performance. Extensive experiments show the efficiency and effectiveness of our FADformer. The source code is available at https://github.com/deng-ai-lab/FADformer.

Keywords: SID · Frequency Learning · Contrastive Regularization

1 Introduction

Single image deraining (SID) is a critical task in low-level image restoration, aiming to restore the clean background image from the rainy one. It is a challenging ill-posed inverse problem due to the unknown true background and rain distribution. Rainy images significantly impact the performance of downstream visual tasks [15, 57], keeping SID at the forefront of research. Early traditional prior-based methods [5,19,25,30] have tried to remove rain by analyzing their statistical characteristics. However, these approaches often fail in cases with dense, complex, and diverse rain streak patterns.

* Corresponding author

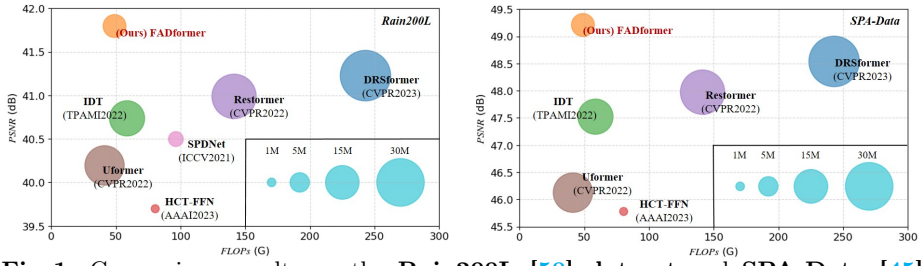

Fig. 1: Comparison results on the **Rain200L [50] dataset** and **SPA-Data [45] dataset**. The circle size reflects the number of model parameters. The proposed method achieves a better trade-off between model complexity and performance.

Recently, learning-based methods [10, 17, 23, 37, 44] have revolutionized image deraining by incorporating diverse CNN architectures, outputting clearer images compared to traditional algorithms. To further improve deraining results, Vision Transformers [2, 4, 46, 49, 52] have been applied to image deraining and achieved remarkable performance as they can better model the long-range information for high-quality image reconstruction. However, this current mainstream supervised learning paradigm still faces two primary challenges: (1) **Lack of efficient global modeling mechanism:** Global modeling capability is often achieved through spatial-domain operations, such as self-attention mechanism [43]. This incurs increasingly high computational costs with mismatched performance improvement as shown in Fig. 1, posing challenges in striking a balance between effectiveness and efficiency. (2) **Insufficient utilization of contrastive samples:** Most methods only consider clear images in the learning process and disregard negative samples as lower-bound information, while fewer methods [13, 55] use a general perceptual-based contrastive regularization which neglects the prior knowledge of deraining tasks, resulting in limited effects.

Unlike most current spatial-domain deraining solutions, we observe some distinct advantages of rainy images from the frequency-domain perspective. Firstly, as shown in Fig. 2.a, the comparison of the first two groups of images illustrates that the different rain streak patterns differ significantly in the frequency domain, and the comparison of the last two groups of images illustrates that the rain degradation patterns are spatially independent of the content in the frequency domain. This demonstrates that Discrete Fourier Transform can extract the density and direction of rain streak noise in the spatial domain and separate it from the image content. Secondly, Fig. 2.b shows that repairing local features in the frequency domain exhibits a globally beneficial influence in the spatial domain, suggesting its capability in modeling extensive long-range information. The combination of these properties helps to establish an efficient long-range capture operation that is sensitive to rain streaks. Furthermore, it reveals the potential of regarding the frequency domain as a contrast space, showcasing significant differences between clear images and those with varied rain streak patterns. In short, the frequency domain provides a new perspective to solve the aforementioned problems.

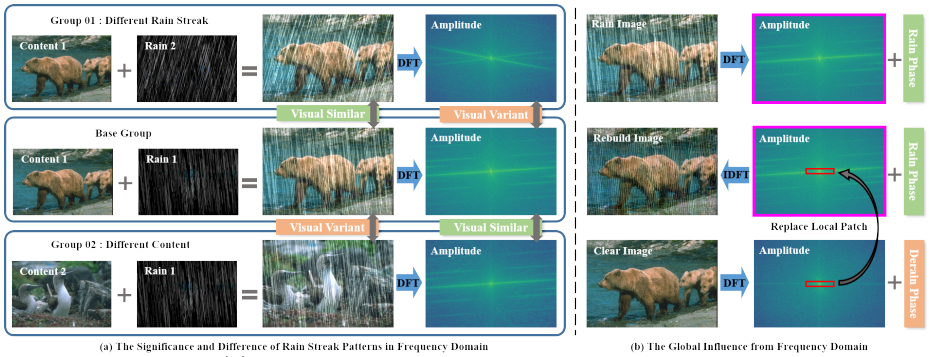

Fig. 2: Motivation. (a) By independently modifying the background content and foreground degraded rain streaks in rainy images, we observe that images with identical rain streaks but different content are relatively similar in the frequency domain, whereas images with different rain streaks but the same content exhibit significant differences. This highlights the Discrete Fourier Transform's (DFT) sensitivity to rain degradation patterns. (b) Replacing local patches in the amplitude components of rain images after DFT has a global impact on the spatial domain after IDFT. This demonstrates the global modeling capacity of frequency domain feature processing.

To address the first challenge, we introduce the **FADformer**, a hierarchical framework designed to balance global-local modeling while maintaining high efficiency. FADformer consists of a series of FADBlocks with different scales, which include two key components: the Fused Fourier Convolution Mixer (**FFCM**) and the Prior-Gated Feed-forward Network (**PGFN**). In practice, FFCM, built upon Fast Fourier Convolution [6] (FFC), blends multi-scale spatial features in the frequency domain to extract global information more efficiently than the vanilla token-mixer in Transformer, which relies on self-attention and often incurs high computational costs. After FFCM, PGFN introduces residue channel prior [22] (RCP) information in a gated manner, guiding the feed-forward network to preserve structural features and enhance local detail restoration capabilities.

Furthermore, to tackle the second problem, we develop a frequency-domain contrastive regularization (**FCR**) term into the loss, which considers the frequency characteristic prior to the image-deraining task. By simply encoding images into the frequency domain, FCR captures salient feature distinctions between predicted outcomes and positive/negative samples, thereby fostering effective contrastive learning. Significant performance improvement on different datasets and models demonstrates the universal validity of FCR.

Our contributions can be summarized as follows:

- We propose an efficient frequency-aware transformer architecture maintaining global-local characteristics to generate high-quality deraining results, which incorporates a frequency-domain convolution mixer and a deraining prior-inspired feed-forward network.
- We introduce a novel frequency-domain contrastive regularization, which conveniently and significantly enhances deraining performance and improves

the utilization of negative sample information. Furthermore, this term is model-agnostic and general to other tasks.
– Comprehensive experiments conducted on real and synthetic datasets show that our approach achieves favorable performance against state-of-the-art methods while keeping an efficient model complexity.

2 Related Works

Deep Single Image Deraining. Early CNN-based models [10,37] achieve better deraining results by modeling rain direction [44] and density [54], surpassing manual feature engineering. Subsequently, techniques like recurrent computations [23] and multi-scale features [17] are introduced, leading to more competitive outcomes. Following this, Transformer-based approaches [46,49,52] in image deraining have gained significant achievement, due to the long-range modeling capability of self-attention mechanisms. Recently, considering that self-attention mechanisms have limitations in repairing local details, some methods [2, 4, 16] enhance Transformer models by improving a convolutional Feed-Forward Network (FFN) or introducing more CNN elements into the Transformer, aiming for better results. However, these models struggle to achieve a balance of efficiency and effectiveness. In contrast to these methods, our approach captures global features via frequency-domain feature processing, enabling the model to balance performance and computational efficiency.

Frequency-domain Features. Recently frequency-domain features have gained significant attention in deep learning and image processing. Fast Fourier Convolution [6] (FFC) leverages the concept of feature learning based on the Fast Fourier Transform for global-to-global feature mapping, delivering impressive performance. In recent image restoration research, Sinha *et al.* [41] introduce non-local FFC for image super-resolution, Qiu *et al.* [36] conduct self-attention over a joint space-time-frequency domain for video super-resolution, Kong *et al.* [20] develop a frequency domain-based self-attention solver for efficient image deblurring, and Zhou *et al.* [56] employ Fourier transformation as an image degradation prior to construct a new global modeling backbone network. In contrast to these methods, we explore the significant characteristics of rain streak noise in the frequency domain and revise the structure of FFC with spatial domain information optimization into the image-deraining architecture, extending its capabilities further.

Contrastive Learning for Image Restoration. Contrastive learning has made significant strides in self-supervised learning tasks in the last few years [1]. Recently, it has found applications in low-level vision tasks, serving as a regularization term to enhance image dehazing [48] and improve performance in unpaired image translation tasks [33], *etc.* In supervised single-image deraining, several works [13, 55] have employed perception-based contrastive regularization and made improvements. In contrast to these methods, we develop an effective frequency-domain-based contrastive regularization that considers rain streak characteristics.

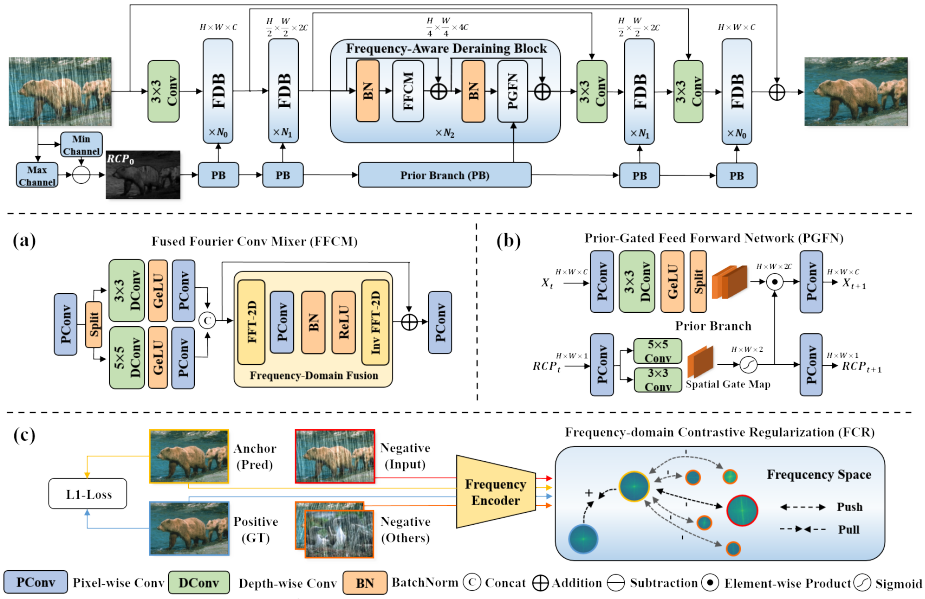

Fig. 3: Illustration of our FADformer framework containing Fused Fourier Convolution Mixer (a, FFCM) and Prior-Gated Feed Forward Network (b, PGFN) with Frequency-domain Contrastive Regularization (c, FCR) for single image deraining.

3 Method

3.1 Overview

Our goals are two-fold: 1) establishing a deraining network that is both efficient and proficient in global-local modeling, and 2) creating a contrastive regularization that utilizes negative samples to enhance deraining performance. The overall workflow of the proposed Frequency-Aware Deraining (FAD) Framework is illustrated in Fig. 3. To achieve the first goal, we construct a hierarchical Transformer-like network called FADformer, which considers frequency-domain information and prior knowledge. As for the second aim, we develop a contrastive regularization term using the frequency domain as the contrast space.

3.2 Backbone Pipeline

Given a rainy image $I \in \mathbf{R}^{H\times W\times 3}$, we first extract shallow features $X_f^0 \in \mathbf{R}^{H\times W\times C}$ using convolution, while simultaneously computing a structural prior map $X_{RCP}^0 \in \mathbf{R}^{H\times W\times 1}$ through Residue Channel Prior (RCP). Subsequently, we pass the features X_f^0 through a three-stage encoder-decoder network for deep feature extraction. Each stage consists of a stack of varying numbers of FADBlocks, possessing distinct spatial resolution domains and channel dimensions to extract multi-scale features. In each FADBlock, unlike the self-attention mechanism, we employ the FFCM for global modeling, involving only convolution

and Fourier Transform operations. Then, the PGFN is utilized to further enrich the features for local details and texture structures with the guidance of X_{RCP}^t, where X_{RCP}^t is obtained from a lightweight Prior Branch (PB) encoding the shallow prior map X_{RCP}^0. The process of downsampling and upsampling features involves pixel-unshuffle and pixel-shuffle operations [39]. Skip connections [38] are employed to fuse encoder and decoder features from the same stage, followed by channel compression through point-wise convolution [12]. Finally, a 3×3 convolution transforms the features into an image space, resulting in the output of residual images $R \in \mathbf{R}^{H \times W \times 3}$. The final reconstructed result is obtained with $Y = I + R$, where Y indicates the output of the FADformer.

3.3 Frequency-Aware Deraining Block

The FADBlock adheres to the paradigm of the Transformer Block [7], enabling long-range feature perception through token mixers, followed by the enhancement of local features through a feed-forward network (FFN). Differently, in the stage of global feature processing, the FADBlock utilizes FFCM to achieve global modeling of spatial domain features through the extraction of frequency domain characteristics, which is more efficient compared to the self-attention mechanism. For local feature processing, guided by prior knowledge of deraining tasks, we introduce PGFN, utilizing the RCP prior as a gate feature to assist in enhancing the structural restoration capabilities of FFN. Mathematically, given feature X_f^{t-1} as the input of the t-th FADBlock, the procedures are expressed as:

$$X_f^{t-\frac{1}{2}} = X_f^{t-1} + FFCM(BN(X_f^{t-1})) \tag{1}$$

$$X_f^t, X_{RCP}^t = X_f^{t-\frac{1}{2}} + PGFN(BN(X_f^{t-\frac{1}{2}}), X_{RCP}^{t-1}) \tag{2}$$

where BN denotes BatchNorm; $X_f^{t-\frac{1}{2}}$ and X_f^t denotes the output of FFCM and PGFN. Besides, X_{RCP}^{t-1} denotes the feature map of RCP, where X_{RCP}^0 is calculated by the rainy image I.

Fused Fourier Convolution Mixer (FFCM). Recalling the Discrete Fourier Transform (DFT), given a feature map $X \in \mathbf{R}^{H \times W \times C}$, DFT can be formalized as follows:

$$\mathcal{F}(u, v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X(h, w) e^{-j2\pi(\frac{hu}{H} + \frac{wv}{W})} \tag{3}$$

where $\mathcal{F}(u, v)$ is based on Fourier space as the complex component; u and v are the coordinates of Fourier space.

As illustrated in Fig. 2 and Eq. (3), the Fourier Transform offers two advantages for image deraining. Firstly, it possesses the ability to separate image degradation components, where rain streak patterns hold salient and invariant characteristics in the frequency domain. Secondly, the transformed frequency components are calculated from all spatial components, naturally acting as a global feature extractor.

Thus, the proposed FFCM leverages these two advantages efficiently and effectively. It benefits from having a similar global receptive field to the self-attention mechanism but with only the computational cost of convolution. The FFCM architecture, as depicted in the Fig. 3.a, employs a spatial-frequency concatenation design to process diverse rain features and achieve local-global feature fusion.

In spatial domain operations, the approach initially utilizes point-wise convolution to elevate the input feature $X_f^{t-1} \in \mathbf{R}^{H\times W\times C}$ and splits it into two groups to extract multi-scale local features, finally acquiring $X_{spatial} \in \mathbf{R}^{H\times W\times 2C}$. Formalized as follows:

$$X_{spatial} = PConv\big(\sigma \cdot Conv_{multi}\big(Split(PConv(X_f^{t-1}))\big)\big) \tag{4}$$

where $Conv_{multi}$ means different kernel size of 3×3 and 5×5 depth-wise convolutions [8] in spatial domain. $PConv$ is the point-wise convolution and σ is a GeLU activation.

In frequency domain operations, $X_{spatial}$ undergoes the DFT to convert it to the real and imaginary components. These concatenated components are subjected to a convolution operation using a 1×1 kernel size. After modulation, the real and imaginary components are split, and the Inverse DFT $\mathcal{F}^{-1}$ converts the frequency domain features back to the spatial domain:

$$X_{\mathcal{R}}, X_{\mathcal{I}} = \mathcal{F}(X_{spatial}) \tag{5}$$

$$\widehat{X_{\mathcal{R}}}, \widehat{X_{\mathcal{I}}} = \sigma \cdot BN(PConv(Concat(X_{\mathcal{R}}, X_{\mathcal{I}}))) \tag{6}$$

$$X_{frequency} = \mathcal{F}^{-1}(\widehat{X_{\mathcal{R}}}, \widehat{X_{\mathcal{I}}}) \tag{7}$$

Finally, the output of FFCM is acquired by a residual structure and PWConv for channel compression:

$$X_f^{t-\frac{1}{2}} = PConv(X_{spatial} + X_{frequency}) \tag{8}$$

FFCM differs in design and application from other recent frequency-domain methods: 1) Unlike FFTformer [20], which combines the *convolution theorem* with self-attention for efficient global modeling, we introduce learnable parameters to extract features in the spectrum directly; 2) Unlike Fourmer [56], which focuses solely on the frequency domain, we use a spatial-frequency concatenation structure for hybrid feature extraction to improve deraining in complex scenes.

Prior-Gated Feed-forward Network (PGFN). Previous works traditionally utilize a standard FFN based on deep features for local feature extraction, neglecting the potential guidance from task-specific deraining priors. In reality, these priors can be incorporated into deep networks and have demonstrated effectiveness. Here, we present the Prior-Gated Feed-forward Network (PGFN) for the first time, integrating prior knowledge into the FFN in a gated manner to improve local details and structural restoration. Specifically, we introduce the residue channel prior [22,51], which effectively preserves clear structures (as

the gray map shown in Fig. 3) by calculating the variance between the maximum and minimum channel components of the rain image, without the need for learning parameters:

$$X^0_{RCP}(h,w) = \max_{c\in r,g,b} I^c(h,w) - \min_{d\in r,g,b} I^d(h,w) \tag{9}$$

As illustrated in Fig. 3.a, PGFN showcases two parallel branches. The main branch follows the structure of DConv Feed-forward Network [24] (DFN). Given feature $X_f^{t-\frac{1}{2}}$, we first broaden the channel dimension via $PConv$ and then refine local details with $DConv$. Meanwhile, the prior branch introduces extended RCP features as group-wise gate weights. This addition significantly fortifies and reconstructs background information disrupted by rain streaks:

$$X_f^t = PConv(\sigma \cdot DConv(PConv(X_f^{t-\frac{1}{2}})) \odot X_{RCP}^{t-\frac{1}{2}}) \tag{10}$$

where $\odot$ is a Hadamard product and $X_{RCP}^{t-\frac{1}{2}}$ is gating maps.

The prior branch is introduced to provide extended RCP features within the Eq. (10). Firstly, the RCP features are processed through $PConv$ to extract high-dimensional features. Group convolutions of different kernel sizes are then applied to generate group-wise gate weights used by DFN. Recognizing the potential interference when directly applying original RCP features to deep layers [51], we employ an iterative encoding approach. This method gradually reduces the channel dimension of the gating map using $PConv$, which is then forwarded as the RCP features output for the next PGFN block, facilitating its deep representation:

$$X_{RCP}^{t-\frac{1}{2}} = Sigmoid(Conv_{group}(PConv(X_{RCP}^{t-1}))) \tag{11}$$

$$X_{RCP}^{t} = \sigma \cdot PConv(X_{RCP}^{t-\frac{1}{2}}) \tag{12}$$

where $Conv_{group}$ means the group convolution with different kernel size.

3.4 Frequency Contrastive Regularization

Loss functions [2,4,18] such as L1/L2 distance and SSIM Loss are widely used in current SID. However, they solely measure the distance between the output and the positive samples, neglecting the inherent information of using negative samples as a lower bound. Inspired by contrastive learning, there are a few contrastive perceptual learning (CPL) methods [13, 55] in currently supervised SID. They use VGG [40] to construct contrastive space features to improve deraining performance while remaining an issue: semantic information primarily relies on clear images and lacks distinctive identification for different rain streak patterns, limiting the effectiveness of CPL in deraining.

To address the problem, we propose a Frequency-domain Contrastive Regularization (FCR) by integrating DFT with contrastive learning, as illustrated in Fig. 3.c. Specifically, FCR employs ground truth as positive samples and rainy

images as negative ones, while using the output of FADformer as an anchor. Leveraging the notable frequency-domain depiction of rain patterns, we utilize DFT for feature extraction instead of a network like VGG, which rarely considers rain streak characteristics. This approach captures the frequency-domain information of images and measures the L1 distance in frequency space. Notably, we use a paired rain image with randomly selected other rain images to form a negative sample group, which provides information on multiple rain patterns. To bring the anchor closer to the positive sample while pushing away negative ones, we construct FCR by measuring the L1 distance between the anchor and pos/neg samples and computing their ratio as follows:

$$L_{FCR} = \frac{1}{n}\sum_{i=1}^{n}\frac{||\mathcal{F}(I_{GT}) - \mathcal{F}(Y)||_1}{||\mathcal{F}(I_i) - \mathcal{F}(Y)||_1} \tag{13}$$

where I_{GT} is the ground truth and n is the number of negative samples.

Contrastive regularization in the frequency domain offers two advantages: (1) the differences between positive and negative samples, as well as among different negative samples, are notably distinct in the frequency domain. This distinctiveness is advantageous for contrastive learning. (2) DFT can be accelerated through the Fast Fourier Transform (FFT) [9], almost without affecting the training speed. Finally, we formulate the overall loss function as follows:

$$L_{Total} = L_{Pixel}(Y, I_{GT}) + \lambda L_{FCR}(Y, I_{GT}, I) \tag{14}$$

where L_{Pixel} means L1Loss and λ is set to 0.1 for balancing loss terms.

4 Experiments and Discussions

4.1 Datasets and Implementation Details

Datasets. We evaluate our framework on seven public benchmarks, including four synthetic datasets (Rain200L/H [50], DDN-Data [10], DID-Data [54]) and three real-world datasets (SPA-Data [45], Internet-Data [45], RE-RAIN [3]). Rain200L and Rain200H both consist of 1800 rainy/clear synthetic pairs for training and 200 ones for testing but with different rain density levels. DDN-Data and DID-Data contain 12600 and 12000 synthetic pairs for training, with the rest of 1400 and 1200 pairs for testing. As for the real-world part, we train our method on large-scale real-world SPA-Data, containing 638492 image pairs for training and 1000 pairs for testing. In addition, we adopt Internet-Data and RE-RAIN to evaluate model generalization, which respectively include 147 and 300 real-world rainy images without ground truth.

Metrics. Following previous works [2,17], we adopt PSNR [14] and SSIM [47] in the Y channel of the YCbCr space for evaluation. For unpaired datasets, we use the non-reference metrics including BRISQUE [32], NIQE [31], and SSEQ [26].

Implementation Details. In our model for comparison, we set the depth of our model to {4,8,10,8,4} and initial channel C to 32. The number of negative

Table 1: Quantitative evaluations on synthetic and real datasets. **Bold** and underline indicate the best and second-best results. FLOPs are tested on a patch size of 256×256.

Method	Synthetic								Real		Overhead	
	Rain200L		Rain200H		DDN-Data		DID-Data		SPA-Data			
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	#Param	FLOPs
(ICCV'15)DSC [30]	27.16	0.8663	14.73	0.3815	27.31	0.8373	24.24	0.8279	34.95	0.9416	-	-
(CVPR'16)GMM [25]	28.66	0.8652	14.50	0.4164	27.55	0.8479	25.81	0.8344	34.30	0.9428	-	-
(CVPR'17)DDN [10]	34.68	0.9671	26.05	0.8059	30.00	0.9041	30.97	0.9116	36.16	0.9457	-	3.75G
(ECCV'18)RESCAN [23]	36.09	0.9697	26.75	0.8353	31.94	0.9345	33.38	0.9417	38.11	0.9707	0.150M	32.12G
(CVPR'19)PReNet [37]	37.80	0.9814	29.04	0.8991	32.60	0.9459	33.17	0.9481	40.16	0.9816	0.169M	66.25G
(CVPR'20)MSPFN [17]	38.58	0.9827	29.36	0.9034	32.99	0.9333	33.72	0.9550	43.43	0.9843	20.89M	595.5G
(CVPR'20)RCDNet [44]	39.17	0.9885	30.24	0.9048	33.04	0.9472	34.08	0.9532	43.36	0.9831	2.958M	194.5G
(CVPR'21)MPRNet [53]	39.47	0.9825	30.67	0.9110	33.10	0.9347	33.99	0.9590	43.64	0.9844	3.637M	548.7G
(AAAI'21)DualGCN [11]	40.73	0.9886	31.15	0.9125	33.01	0.9489	34.37	0.9620	44.18	0.9902	2.73M	-
(ICCV'21)SPDNet [51]	40.50	0.9875	31.28	0.9207	33.15	0.9457	34.57	0.9560	43.20	0.9871	2.982M	96.29G
(CVPR'22)Uformer [46]	40.20	0.9860	30.80	0.9105	33.95	0.9545	35.02	0.9621	46.13	0.9913	20.60M	41.09G
(CVPR'22)Restormer [52]	40.99	0.9890	32.00	0.9329	34.20	0.9571	35.29	0.9641	47.98	0.9921	26.10M	141.0G
(TPAMI'22)IDT [49]	40.74	0.9884	32.10	0.9344	33.84	0.9549	34.89	0.9623	47.35	0.9930	16.39M	58.44G
(AAAI'23)HCT-FFN [4]	39.70	0.9850	31.51	0.9100	33.00	0.9502	33.96	0.9592	45.79	0.9898	0.874M	80.25G
(CVPR'23)DRSformer [2]	41.23	0.9894	32.17	0.9326	34.35	0.9588	35.35	0.9646	48.54	0.9924	33.70M	242.9G
(Ours)FADformer	**41.80**	**0.9906**	**32.48**	**0.9359**	**34.42**	**0.9602**	**35.48**	**0.9657**	**49.21**	**0.9934**	6.958M	48.51G

samples n is set to 2. FADformer is trained using PyTorch [34] on four NVIDIA GeForce RTX 3090 GPUs from scratch for each dataset. During training, we use AdamW [29] optimizer($\beta_1 = 0.9$ and $\beta_2 = 0.999$) with a batch size of 12 and patch size of 256×256 for 150K iterations. The learning rate is initially set to 1×10^{-3} and decreased to 1×10^{-5} by cosine annealing scheme [28]. Training patches are randomly cropped and augmented with rotation and flipping. We evaluate other baselines using their pre-trained models or reliable reported results. If these are unavailable, we retrain the models on the corresponding dataset.

4.2 Comparisons with the State-of-the-arts

Synthetic datasets results. The quantitative results on four synthetic datasets are shown in the first four columns of Tab. 1. Our FADformer outperforms all SOTA methods on synthetic datasets while keeping a better trade-off between computational and memory costs. Notably, FADformer surpasses the concurrent best approach DRSformer by 0.57 dB on Rain200L and 0.35 dB on average in the PSNR term while only costing 20.65% #Param and 19.97% FLOPs. We also compare the qualitative results on synthetic images with different rain streak densities and directions in Fig. 4. CNN-based methods, *e.g.* DualGCN and SPDNet, fail to reconstruct images in dense rain scenes with color distribution shifts. Transformer-based methods with long-range attention perform better but have flaws in local detail repair. For example, DRSformer cannot distinguish background content similar to rain lines in local areas of sparse scenes. Thanks to FFCM and FCR, our method can fuse spatial and frequency domain features while identifying rain streak patterns by extracting negative sample information, thus producing high-quality results in complex rain scenes.

Real-world datasets results. The quantitative results on SPA-Data are shown in the penultimate column of Tab. 1. Our method achieves more outstanding performance in the real-world dataset. We exceed 49dB for the first time on SPA-Data and acquire 0.67dB higher than the SOTA method in the PSNR index. The

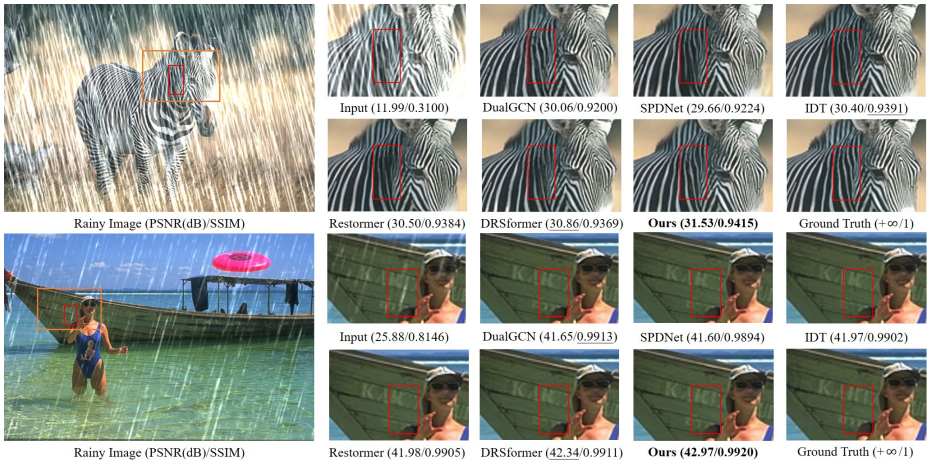

Fig. 4: Qualitative results on synthetic rainy images include the first two rows from the Rain200H and the last two from Rain200L. See supplement for more visualizations.

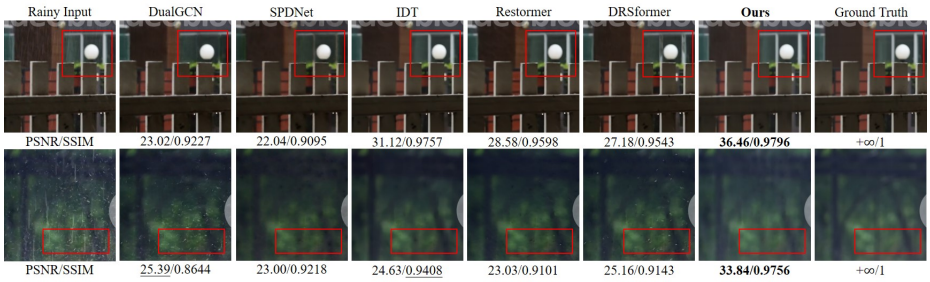

Fig. 5: Qualitative results on real-world paired SPA-dataset.

visual quality comparison is shown in Fig. 5. Our method achieves more pleasant visual effects and can preserve texture details and image structure, which is indispensable to the efficacy of PGFN. In addition, to validate the generalization and effectiveness of FADformer, we evaluate our model and baselines trained with Rain200H on the entire Internet-Data. As shown in Tab. 2, compared with other baselines, FADformer achieves the lowest BRISQUE and SSEQ values, indicating that our results have high-quality content and statistical properties that are closer to natural situations. Qualitative comparison results are shown in Fig. 6. Our results can remove most of the rainlines and have more realistic reconstruction results under different scenes, implying our model can generalize to the real world. More visualizations can be found in supplementary material.

Table 2: Quantitative evaluations on the unpaired full of Internet-Data.

Methods	Rainy Input	MPRNet [53]	SPDNet [51]	Restormer [52]	IDT [49]	HCT-FFN [4]	DRSformer [2]	**Ours**
BRISQUE ↓	28.517	34.733	26.024	32.288	27.042	31.737	26.080	**25.959**
NIQE ↓	5.095	5.144	**4.482**	4.851	4.536	5.131	4.531	4.760
SSEQ ↓	28.280	33.765	27.510	31.789	28.314	30.302	27.954	**26.667**

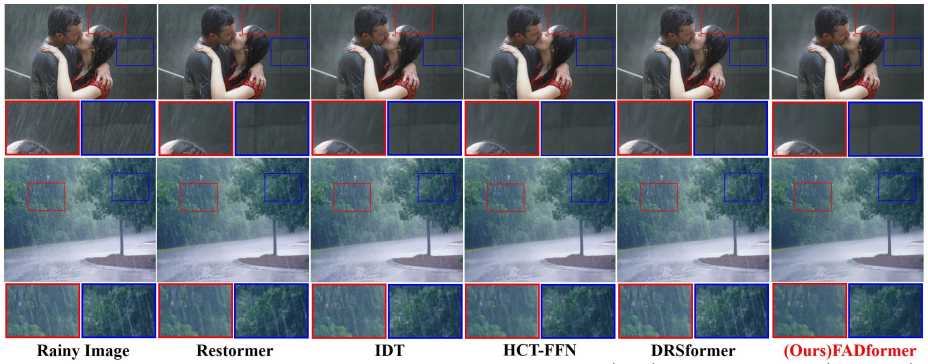

Fig. 6: Visualizing on real-world unpaired Internet-Data(top) and RE-RAIN(bottom).

Other deraining relevant tasks results. We also conduct experiments on raindrop removal, snow removal, and downstream influence tests to further show the deraining effectiveness and generalization ability of our method. Please see the results in the supplementary.

4.3 Ablation Studies

Ablation settings. For consistency and convenience in performance evaluation, our ablation experiments are conducted on Rain200H and adopt a scaled-down version of the experimental setup with a depth of {2,3,4,3,2} and C to 24.
Ablation on FFCM. We first investigate components in FFCM, as shown in Tab. 3. The first two rows prove the efficacy of spatial-domain operations, notably the Multi-Dconv, highlighting the necessity of local feature processing before global feature extraction. The remaining three rows depict the effectiveness of blending global features in the frequency domain. The absence of the Fourier transform operation significantly compromises FFCM's global modeling capacity, leading to a notable reduction. See the supplement for visualizations.

Table 3: Ablation study for key components in FFCM.

Models	Spatial-Domain		Frequency-Domain		Metrics	
	Multi-DConv	PConv	FFT/IFFT	PConv	PSNR	SSIM
baseline(a)		✓	✓	✓	30.13	0.9068
baseline(b)	✓		✓	✓	30.47	0.9094
baseline(c)	✓	✓		✓	29.43	0.8926
baseline(d)	✓	✓			29.27	0.8904
Ours(full FFCM)	✓	✓	✓	✓	**30.73**	**0.9136**

Secondly, we also compare the difference between FFCM and token-mixers based on spatial self-attention in the deraining task. As shown in Tab. 4, we replace FFCM with two different self-attention operations, results show they need more computational burden to acquire similar performance as FFCM by adding network depth. This further proves the efficiency of FADformer.

Table 4: Ablation study for influence of FFCM v.s. Self-attention to FADformer.

Depth Setting		{2,3,4,3,2}	{4,4,8,4,4}	{4,6,10,6,4}	**FADformer (w FFCM)**
FADformer (w Shift-WHSA)	PSNR/SSIM	29.95/0.8980	30.51/0.9060	**30.77**/0.9098	(ablation setting)
(Swin Transformer [27])	FLOPs	8.70G	13.64G	⋆ 16.27G	{2,3,4,3,2}
FADformer (w MDTA)	PSNR/SSIM	29.66/0.8951	30.46/0.9075	30.67/0.9110	30.73/**0.9136**
(Restormer [52])	FLOPs	6.95G	10.56G	⋆ 13.91G	⋆ 12.80G

Ablation on PGFN and Prior Branch. We first compare PGFN with four baselines to evaluate its effectiveness: (1) Feed-forward Network [7], (2) Dconv Feed-forward Network [24], (3) Gated-Dconv Feed-forward Network [52] and (4) Mix-Scale Feed-forward Network [2], as shown in the Tab. 5. Although DFN, GDFN ,and MSFN improve performance by enhancing local perception, they are not effective in deraining tasks without prior information. By adding a prior branch to extract RCP features into DFN, PGFN achieves further improvement.

Table 5: Ablation study for different feed-forward networks.

Models	FN [7]	DFN [24]	GDFN [52]	MSFN [2]	**PGFN**
PSNR/SSIM	29.67/0.9009	30.34/0.9082	30.37/0.9085	30.42/0.9092	**30.73**/**0.9136**

We further explore the components of the Prior Branch in Tab. 6. Results show that all three parts are indispensable. RCP features introduce prior information, Gated Manner can more effectively integrate features in prior and DFN, and Branch Block encodes shallow prior features into deep layers to alleviate problems of feature interference. Visualizations are shown in supplement.

Table 6: Ablation study on individual components of PGFN. w/o ① means that we replace the RCP channel with the mean channel which almost provides no information, and w/o ② means that we use concatenation to introduce the RCP feature map.

Models	①RCP Prior Input	②Gated Manner	③Branch Block	DFN (base)	PSNR	SSIM
baseline(e)				✓	30.34	0.9082
baseline(f)		✓	✓	✓	30.43	0.9086
baseline(g)	✓		✓	✓	30.49	0.9092
baseline(h)	✓	✓		✓	29.52	0.8936
Ours(full PGFN)	✓	✓	✓	✓	**30.73**	**0.9136**

Ablation on FCR. We study the effectiveness of FCR compared with other contrastive regularization terms in Tab. 7. (1) The Contrastive Perceptual Regularization (L1+CPL) using VGG as the encoder has limited effect because it does not consider the rain streak characteristics, while FCR (L1+FCR) takes full advantage of the frequency domain properties of negative samples and achieves obvious advantages. (2) Frequency-domain knowledge (L1+FCR w/o neg, *i.e.* Frequency L1) can lead to some improvements, but there is still a gap compared to FCR with contrastive samples. (3) The negative sample ratio has a great impact on FCR. An appropriate proportion of negative samples can improve the reconstruction effect, while too many negative patterns will cause interference.

Table 7: Ablation study for different contrastive regularization.

Loss Function	L1	L1+CPL (VGG19) [48]	L1+CPL (VGG16) [13]	L1+FCR (w/o neg)	L1+FCR (1:1)	**L1+FCR (1:2)**	L1+FCR (1:4)	L1+FCR (1:6)
PSNR/SSIM	30.25/0.9048	30.42/0.9108	30.53/0.9117	30.48/0.9123	30.62/0.9098	**30.73/0.9136**	30.66/0.9134	30.67/0.9129

4.4 Universal Validity of FCR

We apply FCR to different datasets, methods, and another task to evaluate its universal effectiveness. In Tab. 8, the first three rows show that the incorporation of FCR has significant enhancement effects on different rain removal datasets. The two rows in middle show that FCR can improve the deraining performance of mainstream methods. The last two rows show that FCR has benefits for the learning process of dehazing tasks. Visualizations are shown in the supplement.

Table 8: Results of applying FCR to different scenes. In the methods section, both baselines are retrained on Rain200H instead of using the results in reports. In the tasks section, baselines are retrained on ITS [21] and tested on SOTS-Indoor [21].

Generalization types		w/o FCR		w/ FCR	
		PSNR	SSIM	PSNR	SSIM
Various Datasets	Rain200L	39.66	0.9824	40.21(↑0.55)	0.9869(↑0.0045)
	Rain200H	30.25	0.9048	30.73(↑0.48)	0.9136(↑0.0088)
	DDN-Data	33.51	0.9404	33.61(↑0.10)	0.9413(↑0.0009)
Various Methods	RCDNet [44]	30.10	0.9036	30.35(↑0.25)	0.9054(↑0.0018)
	MPRNet [53]	30.32	0.9049	30.62(↑0.30)	0.9101(↑0.0052)
Other Tasks (Dehazing)	FFA-Net [35]	36.39	0.9890	37.29(↑0.90)	0.9904(↑0.0014)
	Dehazeformer-B [42]	37.84	0.9940	38.40(↑0.56)	0.9947(↑0.0007)

5 Conclusion

In this paper, we introduce a novel frequency-aware deraining framework, called FADformer. By analyzing the issue of inefficient global modeling through the spatial domain, we introduce the Fused Fourier Convolution Mixer to integrate multi-scale spatial features into the frequency domain for efficient global modeling. To enhance structural preservation, we propose the Prior-Gated Feed-forward Network to enhance local detail repair with prior information guidance. Furthermore, to tackle the challenge of leveraging negative sample information, we introduce the FCR, which constructs a contrast space using DFT as an encoder, effectively capturing significant feature differences between positive and negative samples in the frequency domain. Experimental results on synthetic and real-world datasets demonstrate the effectiveness, efficiency, and generalization of the proposed FADformer framework.

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

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
2. 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)
3. Chen, X., Pan, J., Dong, J., Tang, J.: Towards unified deep image deraining: A survey and a new benchmark. arXiv preprint arXiv:2310.03535 (2023)
4. Chen, X., Pan, J., Lu, J., Fan, Z., Li, H.: Hybrid cnn-transformer feature fusion for single image deraining. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 378–386 (2023)
5. Chen, Y.L., Hsu, C.T.: A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In: Proceedings of the IEEE international conference on computer vision. pp. 1968–1975 (2013)
6. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33**, 4479–4488 (2020)
7. 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 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Fran, C., et al.: Deep learning with depth wise separable convolutions. In: IEEE conference on computer vision and pattern recognition (CVPR) (2017)
9. Frigo, M., Johnson, S.G.: Fftw: An adaptive software architecture for the fft. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). vol. 3, pp. 1381–1384. IEEE (1998)
10. 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)
11. Fu, X., Qi, Q., Zha, Z.J., Zhu, Y., Ding, X.: Rain streak removal via dual graph convolutional network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1352–1360 (2021)
12. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
13. Huang, Z., Zhang, J.: Contrastive unfolding deraining network. IEEE Transactions on Neural Networks and Learning Systems (2022)
14. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters **44**(13), 800–801 (2008)
15. Janai, J., Güney, F., Behl, A., Geiger, A., et al.: Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision **12**(1–3), 1–308 (2020)
16. Jiang, K., Wang, Z., Chen, C., Wang, Z., Cui, L., Lin, C.W.: Magic elf: Image deraining meets association learning and transformer. arXiv preprint arXiv:2207.10455 (2022)
17. 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)

18. Jiang, X., Dou, H., Fu, C., Dai, B., Xu, T., Deng, Y.: Boosting supervised dehazing methods via bi-level patch reweighting. In: European Conference on Computer Vision. pp. 57–73. Springer (2022)
19. Kang, L.W., Lin, C.W., Fu, Y.H.: Automatic single-image-based rain streaks removal via image decomposition. IEEE transactions on image processing **21**(4), 1742–1755 (2011)
20. 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)
21. 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)
22. Li, R., Tan, R.T., Cheong, L.F.: Robust optical flow in rainy scenes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 288–304 (2018)
23. 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)
24. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: Localvit: Bringing locality to vision transformers. arXiv preprint arXiv:2104.05707 (2021)
25. Li, Y., Tan, R.T., Guo, X., Lu, J., Brown, M.S.: Rain streak removal using layer priors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2736–2744 (2016)
26. Liu, L., Liu, B., Huang, H., Bovik, A.C.: No-reference image quality assessment based on spatial and spectral entropies. Signal processing: Image communication **29**(8), 856–863 (2014)
27. 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)
28. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
29. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
30. 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)
31. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
32. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
33. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 319–345. Springer (2020)
34. 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)
35. 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)

36. Qiu, Z., Yang, H., Fu, J., Fu, D.: Learning spatiotemporal frequency-transformer for compressed video super-resolution. In: European Conference on Computer Vision. pp. 257–273. Springer (2022)
37. 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)
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
39. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
41. Sinha, A.K., Moorthi, S.M., Dhar, D.: Nl-ffc: Non-local fast fourier convolution for image super resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 467–476 (2022)
42. Song, Y., He, Z., Qian, H., Du, X.: Vision transformers for single image dehazing. IEEE Transactions on Image Processing **32**, 1927–1941 (2023)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
44. 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)
45. 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)
46. 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)
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
48. 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)
49. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.J.: Image de-raining transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
50. 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)
51. 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)
52. 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)
53. 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)
54. 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)
55. Zheng, S., Lu, C., Wu, Y., Gupta, G.: Sapnet: Segmentation-aware progressive network for perceptual contrastive deraining. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 52–62 (2022)
56. 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)
57. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S.: Traffic-sign detection and classification in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2110–2118 (2016)