Learning A Sparse Transformer Network for Effective Image Deraining

Xiang Chen[1]   Hao Li[1]   Mingqiang Li[2]   Jinshan Pan[1*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]Information Science Academy, China Electronics Technology Group Corporation

Abstract

Transformers-based methods have achieved significant performance in image deraining as they can model the non-local information which is vital for high-quality image reconstruction. In this paper, we find that most existing Transformers usually use all similarities of the tokens from the query-key pairs for the feature aggregation. However, if the tokens from the query are different from those of the key, the self-attention values estimated from these tokens also involve in feature aggregation, which accordingly interferes with the clear image restoration. To overcome this problem, we propose an effective DeRaining network, Sparse Transformer (DRSformer) that can adaptively keep the most useful self-attention values for feature aggregation so that the aggregated features better facilitate high-quality image reconstruction. Specifically, we develop a learnable top-k selection operator to adaptively retain the most crucial attention scores from the keys for each query for better feature aggregation. Simultaneously, as the naive feed-forward network in Transformers does not model the multi-scale information that is important for latent clear image restoration, we develop an effective mixed-scale feed-forward network to generate better features for image deraining. To learn an enriched set of hybrid features, which combines local context from CNN operators, we equip our model with mixture of experts feature compensator to present a cooperation refinement deraining scheme. Extensive experimental results on the commonly used benchmarks demonstrate that the proposed method achieves favorable performance against state-of-the-art approaches. The source code and trained models are available at https://github. com/cschenxiang/DRSformer.

1. Introduction

Single image deraining is a typical low-level vision problem emerging in the last decade. It aims to recover the clean
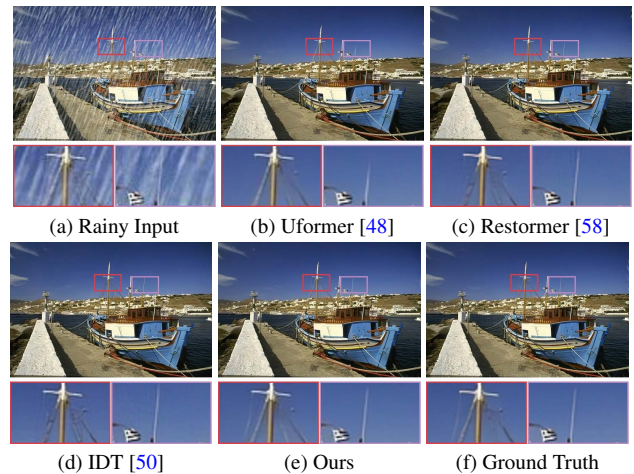
---

*Corresponding author.

Figure 1. Image deraining results between our method and recent Transformer-based methods [48,50,58]. Our method can generate high-quality image with more accurate detail and texture recovery.

image from the observed rainy one. As the clear image and rain streaks are unknown, it is an ill-posed inverse problem. To solve this problem, early approaches [20, 24, 60] usually impose various priors based on statistical properties of rain streaks and clear images. In fact, these handcrafted priors are not robust to complex and varying rainy scenarios, which limit the deraining performance.

Recently, numerous learning-based methods [4, 19, 23, 36, 52, 53, 56] have resorted to diverse CNN architectures as a preferable choice compared to traditional algorithms. However, the intrinsic characteristics of convolutional operation, *i.e*, local receptive fields and independence of input content, hinder the model's capacity to eliminate long-range rain degradation perturbation. To alleviate such limitations, Transformers [2, 26, 35, 50] have been applied to image deraining and have achieved decent performance as they can better model the non-local information for high-quality image reconstruction. Nevertheless, the image details, which are local features of images, are not modeled well by these approaches when restoring clear images as shown in Figure 1. One main reason is that the self-attention in Transformers does not model the local invariant properties that

CNNs do well. Since rain streaks tend to confuse with background details in local regions, recent studies [5, 18, 57] try to mitigate such drawbacks by combining CNN operations and Transformers for boosting image deraining, where the Transformers based on the standard formulations.

We note that the standard Transformers [40] usually use all attention relations based on the query-key pairs to aggregate features. As the tokens from the key are not always relevant to those from the query, using the self-attention values estimated from these tokens in the feature aggregation interferes with the following latent clear image restoration. The root cause behind this deficiency lies in that, the native dense calculation pattern of self-attention amplifies relatively smaller similarity weights, making feature interaction and aggregation process susceptible to implicit noises. This also naturally leads to corresponding redundant or irrelevant representations are still taken into consideration when modeling global feature dependencies [44, 64]. Thus, these findings motivate us to explore the most useful self-attention values so that we can make full use of the features for better image restoration.

To this end, we develop an effective sparse Transformer network for image deraining, named as DRSformer. Specifically, the key component of the proposed framework is the sparse Transformer block (STB) which contains a top-k sparse attention (TKSA) that keeps the most useful self-attention values for feature aggregation and a mixed-scale feed-forward network (MSFN) that explores the multi-scale features for better image deraining. First, we design the top-k attention mechanism to replace the vanilla self-attention [40]. The TKSA keeps the largest K similarity scores between the queries and the keys for the self-attention computing, thereby facilitating better feature aggregation. Furthermore, the developed MSFN further explores the multi-scale information to better improve the aggregated features. Finally, based on the observation that rain distribution reveals the degradation location and degree, we also introduce mixture of experts feature compensator (MEFC) to provide collaborative refinement for STB. With the above-mentioned designs, our proposed method offers three-fold advantages: (1) it can enjoy natural robustness in terms of less sensitivity to useless feature interference, (2) it can not only enrich the locality but also empower the capability of global feature exploitation, and (3) it can co-explore data (embodied in MEFC) and content (embodied in STB) sparsity for achieving deraining performance gains.

The main contributions are summarized as follows:

- We propose a sparse Transformer architecture to help generate high-quality deraining results with more accurate detail and texture recovery.
- We develop a simple yet effective learnable top-k selection operator to adaptively maintain the most useful self-attention values for better feature aggregation.

- We design an effective feed-forward network based on mixed-scale fusion strategy to explore multi-scale representations for better facilitating image deraining.
- Extensive experimental results on various benchmarks demonstrate that our method achieves favorable performance against state-of-the-art (SOTA) approaches.

2. Related Work

Single image deraining. Since image deraining is an ill-posed problem, traditional methods [12, 20, 24, 30, 60] usually develop kinds of image priors to provide additional constraints. However, these handcrafted priors tend to rely on empirical observations and thus are not able to model the inherent properties of clear images. To overcome this problem, numerous CNN-based frameworks [53] have been developed to solve image deraining and achieved decent restoration performance. To better represent the rain distribution, several studies take rain characteristics such as rain direction [27], density [61], veiling effect [15] into account, and optimize the network structure via recursive computation [19, 23, 36] or transfer mechanism [16, 49, 54, 55]. Although these methods achieve better performance than the hand-crafted prior-based ones, they have difficulty capturing the long-range dependencies due to the intrinsic limitations of convolution. Different CNN-based deraining approaches, we utilize the Transformer as the network backbone to model non-local information for image deraining.

Vision Transformers. Motivated by the great success of the Transformers [7] in natural language processing (NLP) [40] and high-level vision tasks [1, 28], Transformers have been applied to image restoration [2, 13, 48, 51, 58] and perform better than the previous CNN-based baselines as they are able to model non-local information. For the field of image rain removal, Jiang *et al.* [18] design a dynamic associated deraining network by incorporating self-attention in Transformer with a background recovery network. More recently, Xiao *et al.* [50] elaborately develop image deraining Transformer (IDT) with window-based and spatial-based dual Transformer to achieve excellent results. Note that, most existing methods rely on the dense dot-product self-attention as the heart of Transformers. However, one shortcoming of this computation manner is that redundant or irrelevant features with smaller weights may interfere with the attention map, which makes the output features contain potential noises. In this work, we propose sparse attention in Transformer to relieve the negligence of the most relevant information faced by vanilla self-attention.

Sparse representation. With inspirations drawn from neural activity in biological brains, sparsity of hidden representation in deep neural networks as a tantalizing "free lunch" emerges for both vision and NLP tasks [44, 64]. Indeed, it is widely proven that sparse representation also plays a critical role in handling low-level vision problems, such as im-
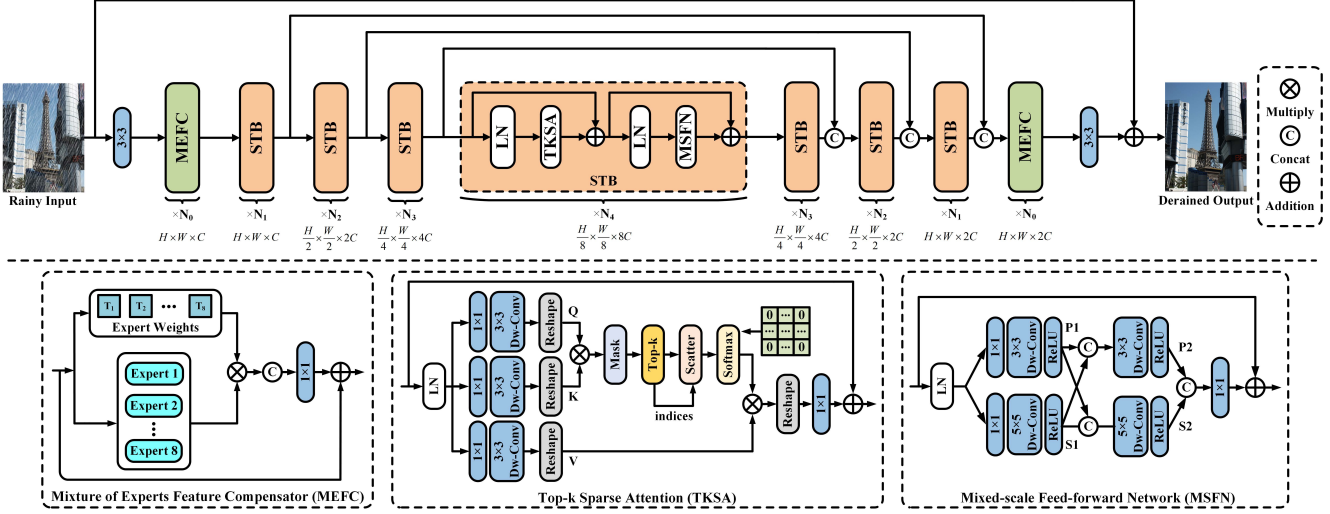
Figure 2. The overall architecture of the proposed sparse Transformer network for image deraining (DRSformer), which mainly contains sparse Transformer block (STB) with top-k sparse attention (TKSA) and mixed-scale feed-forward network (MSFN), and mixture of experts feature compensator (MEFC). LN refers to the layer normalization and DW-Conv refers to the depth-wise convolution.

age deraining [46] and super-resolution [31]. In principle, sparse attention can be categorized into data-based (fixed) sparse attention and content-based sparse attention [6, 38]. For data-based sparse attention, several local attention operations are introduced into CNN backbone, which mainly considers attending only to local window size. Recent studies [11, 42] have investigated enforcing sparsity to Transformer backbone. More recently, Zhang *et al.* [63] design an attention retractable Transformer to allow tokens from sparse areas to interact features, which is data-based sparsity. Different from it, we implement a simple but effective approximation for self-attention based on top-k selection to achieve sparse attention, which is content-based sparsity.

Top-k selection. Zhao *et al.* [64] first propose an explicit selection method based on top-k mechanism in NLP tasks. Driven by their success, k-NN attention [42, 44] is further introduced for boosting vision Transformers. Unlike performing top-k selection in the spatial dimension [44], we design an efficient top-k useful channel selection operator.

3. Proposed Method

In this section, we first describe the overall pipeline and symmetrically hierarchical network architecture for image deraining. Afterward, we provide the details of the proposed sparse Transformer block (STB), as the fundamental building unit of our method, which mainly contains two key elements: top-k sparse attention (TKSA) and mixed-scale feed-forward network (MSFN). Finally, we present the introduced mixture of experts feature compensator (MEFC).

3.1. Overall pipeline

The overall pipeline of our proposed DRSformer, shown in Figure 2, is based on a hierarchical encoder-decoder

framework. Given a rainy image $I_{rain} \in \mathbb{R}^{H \times W \times 3}$, where $H \times W$ represents the spatial resolution of the feature map, we perform overlapped image patch embedding with 3×3 convolution. In the network backbone, we stack $N_{i \in [1,2,3,4]}$ STBs to extract rich features for spatially-varying rain distribution. To excavate the multi-scale representation from rain degeneration, each level of encoder-decoder pipeline covers its own specific spatial resolution and channel dimension. For feature down-sampling and up-sampling, we apply pixel-unshuffle and pixel-shuffle operations. Similar to [48,50,58], we also add skip-connections to bridge across continuous intermediate features for stable training. In each STB, unlike the standard self-attention [7] in Transformer, we develop TKSA to achieve feature sparsity, aiming to enforce the feature aggregation process more effectively. In addition, a MSFN is introduced into STB to enrich multi-scale local information and help image restoration. At the early and final stages of the model learning, we equip our model with N_0 MEFCs to provide complementary feature refinement, so that high-quality clear outputs can be finally reconstructed. With this hybrid formulation, we allow DRSformer to exploit both the adaptive content and the intrinsic property of rainy images, facilitating the separation of undesired rain streaks and latent clear background, and experiments demonstrate that the above design choices yield quality improvements (see Sec. 4.3)

The final reconstructed result is obtained by: $I_{derain} = \mathcal{F}(I_{rain}) + I_{rain}$, where $\mathcal{F}(\cdot)$ is the overall network and it is trained by minimizing the following loss function:

$$\mathcal{L} = \|I_{derain} - I_{gt}\|_1, \tag{1}$$

where I_{gt} denotes the ground-truth image, and $\|\cdot\|_1$ denotes the L_1-norm.

3.2. Sparse Transformer block

As the standard Transformers [7, 40, 58] take all the tokens to compute self-attention globally, which is unfriendly for image restoration due to it may involve noisy interactions between the irrelevant features. To solve such limitations, we develop a sparse Transformer block (STB) as the feature extraction unit by taking the advantages of sparsity that emerged in neural networks [64]. Formally, given the input features at the (l-1)-th block $\mathbf{X}_{l-1}$, the encoding procedures of STB can be defined as:

$$\mathbf{X}'_l = \mathbf{X}_{l-1} + \text{TKSA}\left(\text{LN}\left(\mathbf{X}_{l-1}\right)\right), \quad (2)$$

$$\mathbf{X}_l = \mathbf{X}'_l + \text{MSFN}\left(\text{LN}\left(\mathbf{X}'_l\right)\right), \quad (3)$$

where LN denotes the layer normalization; $\mathbf{X}'_l$ and $\mathbf{X}_l$ denote the outputs from the top-k sparse attention (TKSA) and mixed-scale feed-forward network (MSFN), which are described below.

Top-k sparse attention (TKSA). We revisit the standard self-attention in Transformer, which has become an empirical operation in most of the existing models. Given a query Q, key K and value V with the dimension of $\mathbb{R}^{L \times d}$, the output of dot-product attention is generally formulated as:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\lambda}\right)\mathbf{V}, \quad (4)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the matrix forms of Q, K, and V, respectively. λ is an optional temperature factor defined by $\lambda = \sqrt{d}$. Generally, multi-head attention is implemented to each of the k new Q, K and V, yielding $d = C/k$ channel dimensional outputs which are concatenated and then got the final result for all heads via the linear projection. Noted that, this vanilla self-attention paradigm is based on densely fully-connected, which requires computing the attention map for all query-key pairs. In our work, we develop TKSA to replace it, thus avoiding the involvement of irrelevant information during the feature interaction process.

Specifically, we first encode channel-wise context by applying 1×1 convolutions followed by 3×3 depth-wise convolutions. Inspired by [58], we apply self-attention across channels rather than the spatial dimension to reduce the time and memory complexity. Next, we calculate similarities of pixel pairs between all the reshaped queries and keys, and mask out the unnecessary elements assigned with lower attention weights in the transposed attention matrix $\mathbf{M}$ of size $\mathbb{R}^{\hat{C} \times \hat{C}}$. Unlike the dropout strategy of randomly abandoning the scores, an adaptive selection of the top-k contributive scores is implemented upon $\mathbf{M}$, aiming to preserve the most significant components and remove the useless ones [3]. Here, k is an adjustable parameter to dynamically control the magnitude of sparsity, which is formally obtained by weighted average of some proper fractions, such as $\frac{2}{3}$. Thus, only top-k values within the range $[\Delta_1, \Delta_2]$ are normalized from each row of $\mathbf{M}$ for softmax computing. For

other elements that are smaller than top-k scores, we replace their probabilities with 0 at given indices using the scatter function. This dynamic selection makes the attention from *dense* to *sparse*, which is derived by:

$$\text{SparseAtt}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\mathcal{T}_k(\frac{\mathbf{Q}\mathbf{K}^\top}{\lambda})\right)\mathbf{V}, \quad (5)$$

where $\mathcal{T}_k(\cdot)$ is the learnable top-k selection operator:

$$[\mathcal{T}_k(\boldsymbol{S})]_{ij} = \begin{cases} S_{ij} & S_{ij} \geq t_i \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

where t_i is the k-th largest value in the j-th row of $\frac{\mathbf{Q}\mathbf{K}^\top}{\lambda}$. Here, we calculate $\mathbf{Q}\mathbf{K}^\top$ in the same way as [58] to generate $\mathbf{M}$ of size $\mathbb{R}^{\hat{C} \times \hat{C}}$. As we use the multi-head strategy, we finally concatenate all the outputs of multi-head attention, and then get the final result by the linear projection.

Mixed-scale feed-forward network (MSFN). Previous studies [48, 50, 58] usually introduce single-scale depth-wise convolutions into the regular feed-forward network to imporve locality. However, those exploitations all ignore the correlations of multi-scale rain streaks. In fact, rich multi-scale representation has fully demonstrated its effectiveness [19, 41] in better removing rain. Here, we design a MSFN by inserting two multi-scale depth-wise convolution paths in the transmission process, see Figure 2. Given an input tensor $\mathbf{X}_{l-1} \in \mathbb{R}^{H \times W \times C}$, after layer normalization, we first utilize 1×1 convolution to expand the channel dimension in the ratio of r, then feed it into two parallel branches. During the feature transformation, the 3×3 and 5×5 depth-wise convolutions are employed to enhance the multi-scale local information extraction. In this way, the entire feature fusion procedure of the developed MSFN is formulated as:

$$\hat{\mathbf{X}}_l = f^c_{1 \times 1}\left(\text{LN}\left(\mathbf{X}_{l-1}\right)\right),$$
$$\mathbf{X}_l^{p_1} = \sigma\left(f^{dwc}_{3 \times 3}(\hat{\mathbf{X}}_l)\right), \mathbf{X}_l^{s_1} = \sigma\left(f^{dwc}_{5 \times 5}(\hat{\mathbf{X}}_l)\right),$$
$$\mathbf{X}_l^{p_2} = \sigma\left(f^{dwc}_{3 \times 3}[\mathbf{X}_l^{p_1}, \mathbf{X}_l^{s_1}]\right), \mathbf{X}_l^{s_2} = \sigma\left(f^{dwc}_{5 \times 5}[\mathbf{X}_l^{s_1}, \mathbf{X}_l^{p_1}]\right),$$
$$\mathbf{X}_l = f^c_{1 \times 1}[\mathbf{X}_l^{p_2}, \mathbf{X}_l^{s_2}] + \mathbf{X}_{l-1},$$
$$(7)$$

where $\sigma(\cdot)$ is a ReLU activation, $f^c_{1 \times 1}$ represents 1×1 convolution, $f^{dwc}_{3 \times 3}$ and $f^{dwc}_{5 \times 5}$ denote 3×3 and 5×5 depth-wise convolutions, and $[\cdot]$ is the channel-wise concatenation.

3.3. Mixture of experts feature compensator

To fill in the comprehensive faculty of integrating sparsity in the DRSformer, we further introduce MEFC to perform a unified co-exploration towards joint data and content sparsity. Recalling the classical design of effective CNN models [39], we elaborately select multiple sparse CNN operations to form parallel layers, dubbed as *experts*, which involve an average pooling with receptive field of 3×3, separable convolution layers with kernel sizes of 1×1,

Table 1. Comparison of quantitative results on synthetic and real datasets. **Bold** and underline indicate the best and second-best results.

Datasets		Rain200L		Rain200H		DID-Data		DDN-Data		SPA-Data	
Metrics		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Prior-based methods	DSC [30]	27.16	0.8663	14.73	0.3815	24.24	0.8279	27.31	0.8373	34.95	0.9416
	GMM [24]	28.66	0.8652	14.50	0.4164	25.81	0.8344	27.55	0.8479	34.30	0.9428
CNN-based methods	DDN [8]	34.68	0.9671	26.05	0.8056	30.97	0.9116	30.00	0.9041	36.16	0.9457
	RESCAN [23]	36.09	0.9697	26.75	0.8353	33.38	0.9417	31.94	0.9345	38.11	0.9707
	PReNet [36]	37.80	0.9814	29.04	0.8991	33.17	0.9481	32.60	0.9459	40.16	0.9816
	MSPFN [19]	38.58	0.9827	29.36	0.9034	33.72	0.9550	32.99	0.9333	43.43	0.9843
	RCDNet [43]	39.17	0.9885	30.24	0.9048	34.08	0.9532	33.04	0.9472	43.36	0.9831
	MPRNet [59]	39.47	0.9825	30.67	0.9110	33.99	0.9590	33.10	0.9347	43.64	0.9844
	DualGCN [9]	40.73	0.9886	31.15	0.9125	34.37	0.9620	33.01	0.9489	44.18	0.9902
	SPDNet [56]	40.50	0.9875	31.28	0.9207	34.57	0.9560	33.15	0.9457	43.20	0.9871
Transformer-based methods	Uformer [48]	40.20	0.9860	30.80	0.9105	35.02	0.9621	33.95	0.9545	46.13	0.9913
	Restormer [58]	<u>40.99</u>	<u>0.9890</u>	32.00	<u>0.9329</u>	<u>35.29</u>	<u>0.9641</u>	<u>34.20</u>	<u>0.9571</u>	<u>47.98</u>	0.9921
	IDT [50]	40.74	0.9884	<u>32.10</u>	**0.9344**	34.89	0.9623	33.84	0.9549	47.35	**0.9930**
	DRSformer	**41.23**	**0.9894**	**32.17**	0.9326	**35.35**	**0.9646**	**34.35**	**0.9588**	**48.54**	<u>0.9924</u>

3×3, 5×5, 7×7, and dilated convolution layers with kernel sizes of 3×3, 5×5, 7×7. Different from the conventional mixture of experts [17, 37], our MEFC does not attach an external gating network. Instead, we make the self-attention [14, 21] become a switcher of different experts to adaptively select the importance of diverse representations depending on the inputs. Given an input feature map $\mathbf{X}_{l-1} \in \mathbb{R}^{H \times W \times C}$, we first apply the channel-wise average to generate C-dimensional channel descriptor $\mathbf{z}_c \in \mathbb{R}^C$:

$$\mathbf{z}_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_{l-1}(i,j), \tag{8}$$

where $\mathbf{X}_{l-1}(i,j)$ is the (y,x) position of the feature $\mathbf{X}_{l-1}$. Then, the coefficient vector of each expert is allocated corresponding to the learnable weight matrices $\mathbf{W}_1 \in \mathbb{R}^{T \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{O \times T}$. Here, T is the dimension of the weight matrices. To avoid altering the sizes of its inputs and outputs, we zero pad the input feature maps computed by each expert. Finally, the output of the l-th MEFC is calculated by:

$$\mathbf{T}_l = \mathbf{W}_2 \sigma \left(\mathbf{W}_1 \mathbf{z}_c \right),$$
$$\mathbf{X}_l = f_{1\times 1}^c \left[\sum_{i=1}^{O} f_{exp} \left(\mathbf{X}_l, \mathbf{T}_l \right) \right] + \mathbf{X}_{l-1}, \tag{9}$$

where f_{exp} and O represent the expert operations and the number of experts, respectively. $f_{1\times 1}^c$ represents 1×1 convolution, $\sigma(\cdot)$ is a ReLU function, and $[\cdot]$ is the channel-wise concatenation. With this design, MEFC is now closely linked to the main STBs so that is able to adaptively remove the rainy effects of diverse appearances.

4. Experiments and Analysis

4.1. Experimental settings

Datasets. We implement deraining experiments on multiple public benchmarks, including Rain200L/H [52], DID-Data [61] and DDN-Data [8]. Rain200L and Rain200H contain 1,800 synthetic rainy images for training and 200 ones for

testing. DID-Data and DDN-Data consist of 12,000 and 12,600 synthetic images with different rain directions and density levels. There are 1,200 and 1,400 rainy images for testing. In addition, we also evaluate our method using a large-scale real-world dataset, *i.e*, SPA-Data [45], containing 638,492 image pairs for training and 1,000 testing ones.

Comparison methods. We compare our DRSformer with two prior-based models (DSC [30] and GMM [24]), CNN-based methods (DDN [8], RESCAN [23], PReNet [36], MSPFN [19], RCDNet [43], MPRNet [59], DualGCN [9], and SPDNet [56]), and recent Transformer-based methods (Uformer [48], Restormer [58], and IDT [50]). For recent representative methods (DualGCN, SPDNet, Restormer and IDT), we retrain their models provided by the authors if no pretrained models are provided, otherwise we evaluate them with their online codes for fair comparisons. For other approaches, we refer to some reported results in [10, 50].

Evaluation metrics. We adopt PSNR [34] and SSIM [47] as the evaluation metrics for the above benchmarks. Following previous deraining methods [10, 19], we calculate PSNR and SSIM metrics in Y channel of YCbCr space. For the rainy images without ground truth images, we use the non-reference metrics including NIQE [33] and BRISQUE [32].

Training details. In our model, $\{N_0, N_1, N_2, N_3, N_4\}$ are set to $\{4, 4, 6, 6, 8\}$, and the number of attention heads for four STBs of the same level is set to $\{1, 2, 4, 8\}$. The initial channel C is 48 and the expanding ratio is set to 2. In terms of MEFC, we set $O = 8$ for the number of experts and $T = 32$ for the weight matrix. Note that we do not use MEFC for training Rain200L and SPA-Data, because their rain streaks are less complex and easier to learn. In terms of STB, the sparseness $[\Delta_1, \Delta_2]$ in TKSA is set to $[\frac{1}{2}, \frac{4}{5}]$, and the channel expansion factor r in MSFN is set to 2.66. During training, we use AdamW optimizer with batch size of 8 and patch size of 128 for total 300K iterations. The initial learning rate is fixed as 1×10^{-4} for 92K iterations, and then decreased to 1×10^{-6} for 208K iterations with the cosine annealing scheme [29]. For data augmentation, vertical and
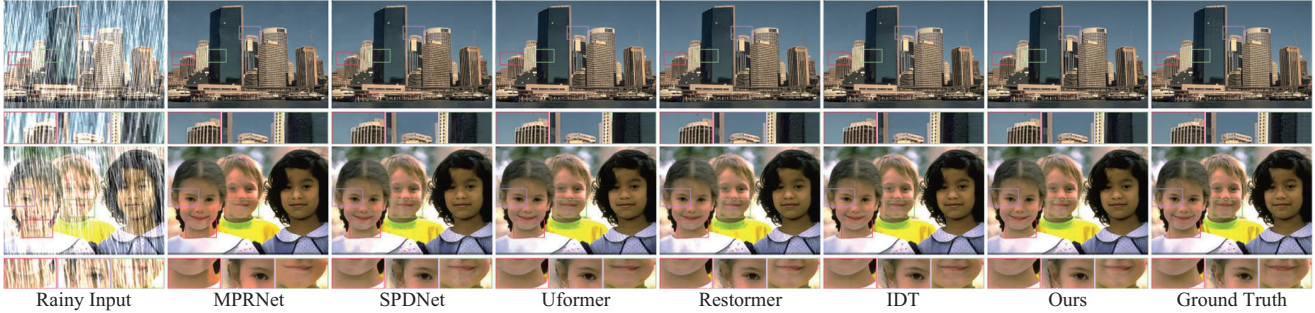
Figure 3. Visual quality comparison on the Rain200H dataset. Zooming in the figures offers a better view at the deraining capability.
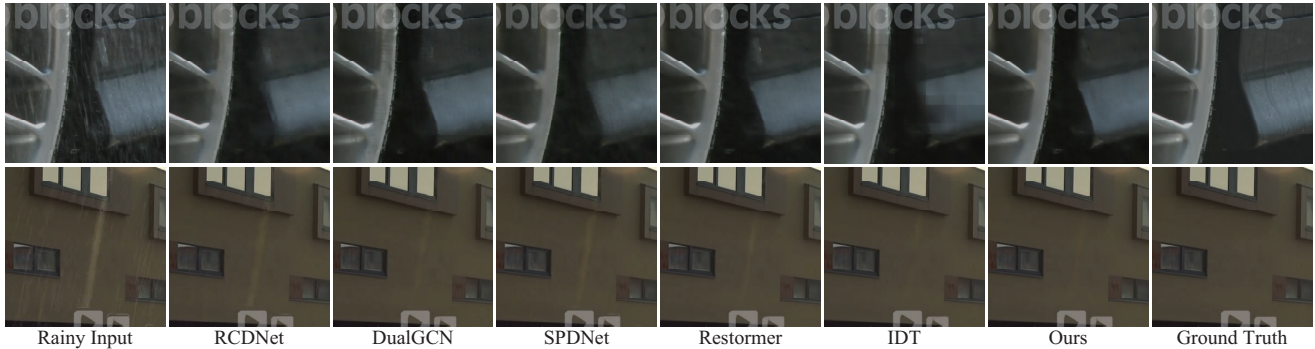

Figure 4. Visual quality comparison on the SPA-Data dataset. Zooming in the figures offers a better view at the deraining capability.

horizontal flips are randomly applied. The entire framework is performed on the PyTorch with 4 NVIDIA GeForce RTX 3090 GPUs, which works in an end-to-end learning fashion without costly large-scale pretraining [2].

4.2. Comparisons with the state-of-the-arts

Synthetic datasets. The quantitative evaluations on different benchmark datasets are reported in Table 1. As shown, we can note that our proposed method outperforms all the other derainers, especially on PSNR, *e.g.*, DRSformer surpasses the concurrent approach IDT by 0.4 dB on average. Compared with previous CNN-based models, this progress can be much more obvious. The notable increasing scores on the DID-Data and DDN-Data benchmarks reveal that our method can properly handle diverse types of spatially-varying rain streaks. For convincing evidence, we show the visual quality comparison between samples generated by recent approaches in Figure 3. The pure CNN-based models, *e.g.*, MPRNet and SPDNet, fail to restore clear images in heavy rainy scenarios. It can be seen that the results of all computing Transformer-based methods are flawed in terms of detail and texture recovery. Unfortunately, IDT even introduces considerable boundary artifacts. Thanks to the developed sparse attention with top-k selection, our method can generate high-quality deraining results, which are more consistent with that of the ground truth.

Real-world datasets. We further conduct experiments on the SPA-Data benchmark dataset, and corresponding results are provided in the last column of Table 1. As expected, our model continues to achieve the highest PSNR/SSIM value, exhibiting the superior of DRSformer in terms of deraining performance and generalization. The visual quality comparison can be observed in Figure 4. In contrast, our method significantly competes with others in removing most rain streaks while preserving truthful image structures. In order to further validate the effectiveness of DRSformer, we also randomly choose 20 real rainy images without ground truths from Internet-Data [45] to perform another evaluation. As displayed in Table 2, our net gets the lower NIQE and BRISQUE values, which means a high-quality output with clearer content and better perceptual quality against other comparative models on the real rainy scenarios. Through qualitative quality comparison in Figure 5, most deep models are sensitive to spatially-long rain streaks and leave some apparent rain effects. On the contrary, our net successfully removes most rain perturbation and owns a visual pleasant recovery effect, which implies that it can well generalize to unseen real-world data types.

4.3. Ablation studies

Effectiveness of Top-k selection. To examine the effect of the top-k selection in the TKSA, we present the deraining results of TKSA w/o top-k in Figure 6. We can see that the PSNR values of the derained images by the method without using the top-k selection are lower than those by the method using the top-k selection. In addition, we also note that each element of the self-attention matrix $\text{softmax}\left(\frac{\mathbf{QK}^\top}{\lambda}\right)$ in Eq. (4) is non-negative and the summation of the elements from each row of $\text{softmax}\left(\frac{\mathbf{QK}^\top}{\lambda}\right)$ is 1. Thus, applying

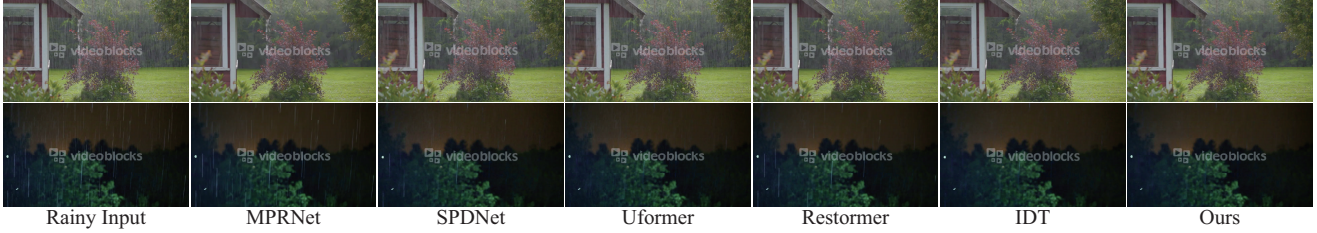| Rainy Input | MPRNet | SPDNet | Uformer | Restormer | IDT | Ours |

Figure 5. Visual quality comparison on real-world rainy images. Zooming in the figures offers a better view at the deraining capability.

Table 2. Comparison of quantitative results on real-world rainy images, and note that lower scores indicate better image quality.

Methods	Rainy Input	MPRNet [59]	SPDNet [56]	Uformer [48]	Restormer [58]	IDT [50]	Ours
NIQE ↓ / BRISQUE ↓	5.829 / 33.129	4.740 / 32.018	4.422 / 26.173	4.833 / 28.106	5.005 / 34.036	4.238 / 25.573	**4.095 / 22.730**

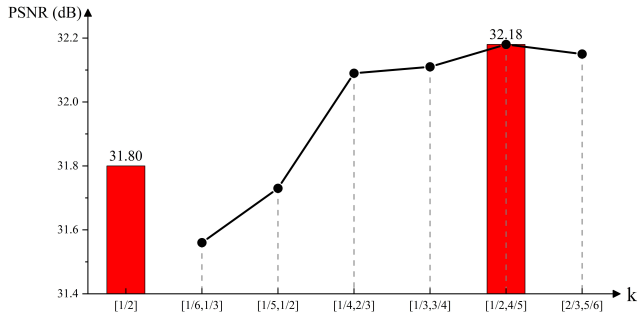Figure 6. Ablation analysis for top-*k* selection on the benchmarks.

Figure 7. Ablation analysis for different number *k* in the TKSA.

| (a) w/o Top-k | (a) w Top-k | (b) w/o Top-k | (b) w Top-k |

Figure 8. Visualization of feature maps. Our proposed top-*k* selection can effectively leverage pixel-dependent properties of image structure and generate more precise high-frequency details.

Effect of the number of *k*. The key parameter for our proposed TKSA is *k*, and its influence is investigated in Figure 7. We note that the optimal choice of *k* determines the boundary control of the sparsity rate. If *k* is manually set to a single value, such as $\frac{1}{2}$, we notice that its performance is sensitive to *k*. To avoid an exhaustive search, we set a controllable interval range for *k* to dynamically learn the most contribute score. When *k* is too small, we find that the performance will undoubtedly decline sharply due to insufficient global information aggregation. The best result can achieve 32.18 dB when $[\Delta_1, \Delta_2]$ in the TKSA is assigned to $[\frac{1}{2}, \frac{4}{5}]$. As *k* continues to increase, the final deraining performance gradually decreases due to the introduction of irrelevant and useless features.

Effectiveness of MSFN. To evaluate the effectiveness of the proposed MSFN, we compare it with three baselines: (1) conventional feed-forward network (FN) [7], (2) Dconv feed-forward network (DFN) [25], and (3) gated-Dconv feed-forward network (GDFN) [58]. The quantitative analysis results on Rain200H are listed in Table 3. Although GDFN introduces a gating mechanism in two same-scale depth-wise convolution streams to bring performance advantages, it still neglects the multi-scale knowledge for deraining. By adding local feature extraction and fusion at different scales, the MSFN can indeed better boost the performance, and achieve PSNR gain of 0.21 dB over GDFN.

Effectiveness of MEFC. To evaluate the effectiveness of MEFC, we perform experiments based on different model variants in Table 4. Compared to the baseline model (a), MEFC provides additional performance benefits thanks to auxiliary data sparsity. In addition, we observe that MEFCs

the self-attention matrix $\mathrm{softmax}\left(\frac{\mathbf{QK}^\top}{\lambda}\right)$ to $\mathbf{V}$ will remove the high-frequency information of $\mathbf{V}$, which lead to over-smoothed results.

To understand the effect of such top-*k* selection, we further use high-pass filtering (HPF) to visualize learned features in Figure 8. Compared to standard self-attention operation (w/o top-*k*), our strategy can better help reconstruct finer-detail feature and improve potential restoration quality. As the nearby pixels tend to be more similar than others, top-*k* selection operator helps to reduce the irrelevant context from long-range pixel dependency. This step of selection allows the smaller similarity weights (from a part of long-range feature interactions) to be discarded in the procedure of self-attention calculation, thus facilitating more accurate representation for achieving high-quality output.
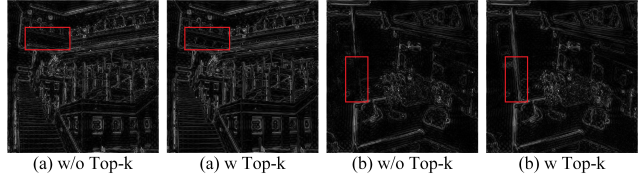
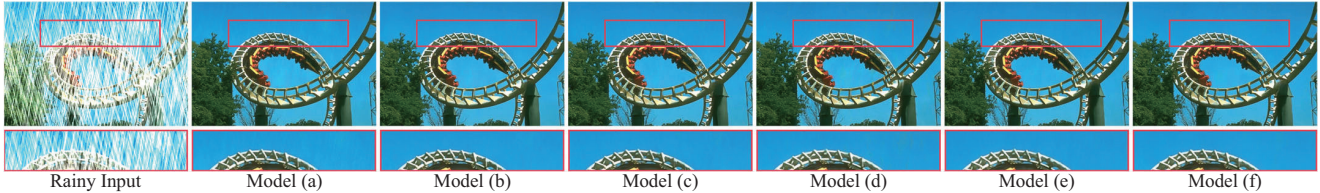| Rainy Input | Model (a) | Model (b) | Model (c) | Model (d) | Model (e) | Model (f) |

Figure 9. Ablation qualitative comparison for different variants of DRSformer. The models (a-f) are consistent with the settings in Table 4.

Table 3. Ablation study for different feed-forward networks.

Models	FN [7]	DFN [25]	GDFN [58]	MSFN
PSNR / SSIM	31.84 / 0.9275	31.88 / 0.9279	31.97 / 0.9286	**32.18 / 0.9330**

Table 4. Ablation study for different variants of our DRSformer. MEFC-1 and MEFC-2 denote MEFC in early and final stages.

Models	MEFC-1	STBs	MEFC-2	Experts	PSNR / SSIM
(a)		✓		0	32.03 / 0.9308
(b)	✓	✓		8	32.01 / 0.9311
(c)		✓	✓	8	32.07 / 0.9328
(d)	✓	✓	✓	1	32.06 / 0.9316
(e)	✓	✓	✓	4	32.14 / 0.9325
(f)	✓	✓	✓	8	**32.18 / 0.9330**

at different locations of the network pipeline have specific impacts on the restoration performance. Indeed, we also analyze the effect of the different numbers of experts in each MEFC. When using single expert model (d), the performance is dramatically degraded compared with our multi-expert model (f). Unlike setting all experts to the same structure [21], our multi-expert formulation is more diverse, which brings its gains to the performance due to different receptive fields and disparate CNN operations. Through the zoomed boxes in Figure 9, the recovered results of the model with all the above components tend to be clearer since it enables more diverse features to be fully used during the restoration process. All in all, our model (f) performs better than the other possible configurations, which indicates that each design strategy that we consider has its own contribution to the final performance of DRSformer.

4.4. Closely-related methods

We note that the recent method [22] proposes a k-NN image Transformer (KiT) to solve image restoration by aggregating k similar patches with the pair-wise local attention. Compared with KiT that employs complex locality sensitive hashing that cannot ensure sufficient global interaction, our simple but effective top-k selection mechanism not only enjoys the locality but also empowers the ability of global relation mining. As the code of KiT is not available, we refer to the results of their paper. Figure 10 shows qualitative comparisons trained on the Rain800 [62]. We can see that KiT tends to blur the contents and cause color distortion. In contrast, our method leads to better deraining results.

In addition, we also note that [44] recently designs the k-NN attention to enhance the representation ability of vision Transformers by selecting the top-k similar tokens. Different from KVT [44], which implements top-k selection

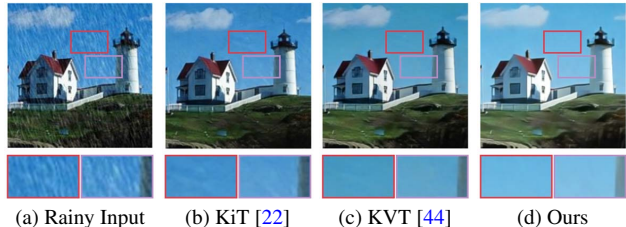| (a) Rainy Input | (b) KiT [22] | (c) KVT [44] | (d) Ours |

Figure 10. Comparison results with closely-related methods.

in the spatial dimension, our operator is more efficient in computing sparse attention across channels. Furthermore, the sparsity level of k in our proposed TKSA is dynamically learnable, rather than the fixed setting in [44]. Here, we adopt k-NN attention in KVT to replace our TKSA for comparison. To ensure the fair comparison, the same training settings are kept for model testing. As shown in Figure 10 (c) and (d), our method can generate a clearer image.

5. Concluding Remarks

We have presented an effective sparse Transformer network, called DRSformer, to solve image deraining. Based on the observation that vanilla self-attention in Transformer may suffer from the global interaction of irrelevant information, we develop the top-k sparse attention to keep the most useful self-attention values for better feature aggregation. To facilitate the aggregated features for removing rain, we develop a mixed-scale feed-forward network to better explore multi-scale representations. Furthermore, the mixture of experts feature compensator is introduced to the model to provide collaborative refinement for the sparse Transformer backbone, so that the fine details of the reconstructed image is preserved. Experimental results show that our DRSformer performs favorably against state-of-the-art methods.

Limitations. Our proposed method aims to further boost image deraining performance, but there are limitations in the model efficiency. Specifically, our model requires 33.7 Million parameters and costs 242.9G FLOPs on one image with size of 256×256. We will apply the pruning or distillation scheme in our model to maintain the original deraining performance while achieving credible model compression.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 1, 2, 6

[3] Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems*, 34:19974–19988, 2021. 4

[4] Xiang Chen, Jinshan Pan, Kui Jiang, Yufeng Li, Yufeng Huang, Caihua Kong, Longgang Dai, and Zhentao Fan. Unpaired deep image deraining using dual contrastive learning. In *CVPR*, pages 2017–2026, 2022. 1

[5] Xiang Chen, Jinshan Pan, Jiyang Lu, Zhentao Fan, and Hao Li. Hybrid cnn-transformer feature fusion for single image deraining. In *AAAI*, 2023. 2

[6] Gonçalo M Correia, Vlad Niculae, and André FT Martins. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3, 4, 7, 8

[8] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, pages 3855–3863, 2017. 5

[9] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding. Rain streak removal via dual graph convolutional network. In *AAAI*, pages 1352–1360, 2021. 5

[10] Xueyang Fu, Jie Xiao, Yurui Zhu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Continual image deraining with hypergraph convolutional networks. *IEEE TPAMI*, 2023. 5

[11] Zhihong Fu, Zehua Fu, Qingjie Liu, Wenrui Cai, and Yunhong Wang. Sparsett: Visual tracking with sparse transformers. *arXiv preprint arXiv:2205.03776*, 2022. 3

[12] Shuhang Gu, Deyu Meng, Wangmeng Zuo, and Lei Zhang. Joint convolutional analysis and synthesis sparse representation for single image layer separation. In *ICCV*, 2017. 2

[13] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, pages 5812–5820, 2022. 2

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 5

[15] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *CVPR*, pages 8022–8031, 2019. 2

[16] Huaibo Huang, Aijing Yu, and Ran He. Memory oriented transfer learning for semi-supervised image deraining. In *CVPR*, pages 7732–7741, 2021. 2

[17] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 5

[18] Kui Jiang, Zhongyuan Wang, Chen Chen, Zheng Wang, Laizhong Cui, and Chia-Wen Lin. Magic elf: Image deraining meets association learning and transformer. In *ACM MM*, 2022. 2

[19] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, pages 8346–8355, 2020. 1, 2, 4, 5

[20] Li-Wei Kang, Chia-Wen Lin, and Yu-Hsiang Fu. Automatic single-image-based rain streaks removal via image decomposition. *IEEE TIP*, 21(4):1742–1755, 2011. 1, 2

[21] Sijin Kim, Namhyuk Ahn, and Kyung-Ah Sohn. Restoring spatially-heterogeneous distortions using mixture of experts network. In *ACCV*, 2020. 5, 8

[22] Hunsang Lee, Hyesong Choi, Kwanghoon Sohn, and Dongbo Min. Knn local attention for image restoration. In *CVPR*, pages 2139–2149, 2022. 8

[23] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, pages 254–269, 2018. 1, 2, 5

[24] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *CVPR*, pages 2736–2744, 2016. 1, 2, 5

[25] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 7, 8

[26] Yuanchu Liang, Saeed Anwar, and Yang Liu. Drt: A lightweight single image deraining recursive transformer. In *CVPRW*, pages 589–598, 2022. 1

[27] Yang Liu, Ziyu Yue, Jinshan Pan, and Zhixun Su. Unpaired learning for deep image deraining with rain direction regularizer. In *ICCV*, pages 4753–4761, 2021. 2

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2016. 5

[30] Yu Luo, Yong Xu, and Hui Ji. Removing rain from a single image via discriminative sparse coding. In *ICCV*, pages 3397–3405, 2015. 2, 5

[31] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *CVPR*, pages 3517–3526, 2021. 3

[32] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012. 5

[33] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE SPL*, 20(3):209–212, 2012. 5

[34] Huynh-Thu Q. and Ghanbari M. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008. 5

[35] Qin Qin, Jingke Yan, Qin Wang, Xin Wang, Minyao Li, and Yuqing Wang. Etdnet: An efficient transformer deraining model. *IEEE Access*, 9:119881–119893, 2021. 1

[36] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, pages 3937–3946, 2019. 1, 2, 5

[37] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, pages 3253–3261, 2018. 5

[38] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. 3

[39] Masanori Suganuma, Xing Liu, and Takayuki Okatani. Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. In *CVPR*, pages 9039–9048, 2019. 4

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[41] Cong Wang, Xiaoying Xing, Yutong Wu, Zhixun Su, and Junyang Chen. Dcsfn: Deep cross-scale fusion network for single image rain removal. In *ACM MM*, pages 1643–1651, 2020. 4

[42] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *CVPR*, pages 7297–7307, 2022. 3

[43] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *CVPR*, pages 3103–3112, 2020. 5

[44] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Wen Xie, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. In *ECCV*, 2022. 2, 3, 8

[45] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, pages 12270–12279, 2019. 5, 6

[46] Yinglong Wang, Chao Ma, and Bing Zeng. Multi-decoding deraining network and quasi-sparsity based training. In *CVPR*, pages 13375–13384, 2021. 3

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[48] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 1, 2, 3, 4, 5, 7

[49] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *CVPR*, pages 3877–3886, 2019. 2

[50] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE TPAMI*, 2022. 1, 2, 3, 4, 5, 7

[51] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 2

[52] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, pages 1357–1366, 2017. 1, 5

[53] Wenhan Yang, Robby T Tan, Shiqi Wang, Yuming Fang, and Jiaying Liu. Single image deraining: From model-based to data-driven and beyond. *IEEE TPAMI*, 2020. 1, 2

[54] Rajeev Yasarla, Vishwanath A Sindagi, and Vishal M Patel. Syn2real transfer learning for image deraining using gaussian processes. In *CVPR*, pages 2726–2736, 2020. 2

[55] Yuntong Ye, Yi Chang, Hanyu Zhou, and Luxin Yan. Closing the loop: Joint rain generation and removal via disentangled image translation. In *CVPR*, pages 2053–2062, 2021. 2

[56] Qiaosi Yi, Juncheng Li, Qinyan Dai, Faming Fang, Guixu Zhang, and Tieyong Zeng. Structure-preserving deraining with residue channel prior guidance. In *ICCV*, pages 4238–4247, 2021. 1, 5, 7

[57] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *ICCV*, pages 579–588, 2021. 2

[58] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 1, 2, 3, 4, 5, 7, 8

[59] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021. 5, 7

[60] He Zhang and Vishal M Patel. Convolutional sparse and low-rank coding-based rain streak removal. In *WACV*, pages 1259–1267, 2017. 1, 2

[61] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, pages 695–704, 2018. 2, 5

[62] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE TCSVT*, 30(11):3943–3956, 2019. 8

[63] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *ICLR*, 2023. 3

[64] Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. Explicit sparse transformer: Concentrated attention through explicit selection. *ICLR*, 2020. 2, 3, 4