Learning A Sparse Transformer Network for Effective Image Deraining

Xiang Chen¹ Hao Li³ Mingqiang Li² Jinshan Pan¹*
¹School of Computer Science and Engineering, Nanjing University of Science and Technology
²Information 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 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 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] elaborate 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 $I_{\text{rain}} \in \mathbb{R}^{H \times W \times 3}$, where $H \times W$ represents the spatial resolution of the feature map, we perform overlapped image patch embedding with $3 \times 3$ convolution. In the network backbone, we stack $N \in \{1, 2, 3, 4\}$ 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: $I_{\text{derain}} = F(I_{\text{rain}}) + I_{\text{rain}}$, where $F(\cdot)$ is the overall network and it is trained by minimizing the following loss function:

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

where $I_{\text{gt}}$ 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 \(X_{l-1}\), the encoding procedures of STB can be defined as:

\[
X'_l = X_{l-1} + \text{TKSA (LN}(X_{l-1})) ,
\]

\[
X_l = X'_l + \text{MSFN (LN}(X'_l)),
\]

where LN denotes the layer normalization; \(X'_l\) and \(X_l\) denote the outputs from the \((l-1)\)-th 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:

\[
\text{Att}(Q, K, V) = \text{softmax} \left( \frac{QK^\top}{\lambda} \right) V,
\]

where \(Q\), \(K\), and \(V\) denote the matrix forms of \(Q\), \(K\), and \(V\), respectively. \(\lambda\) 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 \(1 \times 1\) convolutions followed by \(3 \times 3\) 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 \(M\) of size \(\mathbb{R}^{C \times C}\). Unlike the dropout strategy of randomly abandoning the scores, an adaptive selection of the top-\(k\) contributive scores is implemented upon \(M\), 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 \(M\) 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:

\[
\text{SparseAtt}(Q, K, V) = \text{softmax} \left( \frac{QK^\top}{\lambda} \right) V,
\]

where \(T_k(\cdot)\) is the learnable top-\(k\) selection operator:

\[
[T_k(S)]_{ij} = \begin{cases} S_{ij} & S_{ij} \geq t_i \\ 0 & \text{otherwise} \end{cases}
\]

where \(t_i\) is the \(k\)-th largest value in the \(j\)-th row of \(QK^\top\). Here, we calculate \(QK^\top\) in the same way as [58] to generate \(M\) of size \(\mathbb{R}^{C \times C}\). 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 improve 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 \(X_{l-1} \in \mathbb{R}^{H \times W \times C}\), after layer normalization, we first utilize \(1 \times 1\) convolution to expand the channel dimension in the ratio of \(r\), then feed it into two parallel branches. During the feature transformation, the \(3 \times 3\) and \(5 \times 5\) 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:

\[
\hat{X}_l = f_{c \times 1}^c (\text{LN}(X_{l-1})) ,
\]

\[
X_{l1}^p = \sigma \left( f_{dwc}^{3 \times 3}(\hat{X}_l) \right),
\]

\[
X_{l1}^s = \sigma \left( f_{dwc}^{j \times 5}(\hat{X}_l) \right),
\]

\[
X_{l2}^p = \sigma \left( f_{dwc}^{3 \times 3}[X_{l1}^p, X_{l1}^s] \right),
\]

\[
X_{l2}^s = \sigma \left( f_{dwc}^{j \times 5}[X_{l1}^p, X_{l1}^s] \right),
\]

\[
X_l = f_{c \times 1}^{c + 2}(X_{l1}^p, X_{l2}^s) + X_{l-1},
\]

where \(\sigma(\cdot)\) is a ReLU activation, \(f_{c \times 1}^{c}\) represents \(1 \times 1\) convolution, \(f_{dwc}^{3 \times 3}\) and \(f_{dwc}^{j \times 5}\) denote \(3 \times 3\) and \(5 \times 5\) 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 \(3 \times 3\), separable convolution layers with kernel sizes of \(1 \times 1\),
3 \times 3, 5 \times 5, 7 \times 7$, and dilated convolution layers with kernel sizes of $3 \times 3, 5 \times 5, 7 \times 7$. Different from the conventional mixture of experts [17, 37], our MEF 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 $X_{l-1} \in \mathbb{R}^{H \times W \times C}$, we first apply the channel-wise average to generate $C$-dimensional channel descriptor $z_c \in \mathbb{R}^C$:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{l-1}(i, j),$$  

(8)

where $X_{l-1}(i, j)$ is the $(y, x)$ position of the feature $X_{l-1}$. Then, the coefficient vector of each expert is allocated corresponding to the learnable weight matrices $W_1 \in \mathbb{R}^{T \times C}$ and $W_2 \in \mathbb{R}^{O \times T}$. 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 MEF is calculated by:

$$T_l = W_2 \sigma(W_1 z_c),$$

$$X_l = f_{\times 1x1} \left[ \sum_{i=1}^{O} f_{exp} (X_i, T_l) \right] + X_{l-1},$$  

(9)

where $f_{exp}$ and $O$ represent the expert operations and the number of experts, respectively. $f_{\times 1x1}$ represents $1 \times 1$ convolution, $\sigma(\cdot)$ is a ReLU function, and $[\cdot]$ is the channel-wise concatenation. With this design, MEF 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 training. 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], MSFNN [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 MEF, we set $O = 8$ for the number of experts and $T = 32$ for the weight matrix. Note that we do not use MEF 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 $[\frac{1}{4}, \frac{1}{2}]$, and the channel expansion factor $r$ in MSFNN 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 $1 \times 10^{-4}$ for 92K iterations, and then decreased to $1 \times 10^{-6}$ for 208K iterations with the cosine annealing scheme [29]. For data augmentation, vertical and
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 \( \text{softmax} \left( \frac{QK^T}{\lambda} \right) \) in Eq. (4) is non-negative and the summation of the elements from each row of \( \text{softmax} \left( \frac{QK^T}{\lambda} \right) \) is 1. Thus, applying
the self-attention matrix $\text{softmax} \left( \frac{QK^\top}{\lambda} \right)$ to $V$ will remove the high-frequency information of $V$, 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.

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 $[\frac{1}{2}, \frac{1}{2}]$. 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 de-raining. 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
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 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 $256 \times 256$. We will apply the pruning or distillation scheme in our model to maintain the original deraining performance while achieving credible model compression.

Acknowledgements. This work has been partly supported by the National Key R&D Program of China (No. 2018AAA0102001), the National Natural Science Foundation of China (Nos. U22B2049, U19B2040, 61922043, 61872421, 62272230), and the Fundamental Research Funds for the Central Universities (No. 30920041109).
References


