Dilated Convolutional Transformer for High-Quality Image Deraining

Yufeng Li[1] Jiyang Lu[1] Hongming Chen[1] Xianhao Wu[1] Xiang Chen[2*]
[1]College of Electronic and Information Engineering, Shenyang Aerospace University
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology

Abstract

Convolutional neural networks (CNNs) and Transformers have achieved significant success in image signal processing. However, little effort has been made to effectively combine the properties of these two architectures to satisfy image deraining. In this paper, we propose an effective deraining method, dilated convolutional Transformer (DCT), which can enlarge the receptive fields of the network to aggregate global information. The fundamental building unit of our approach is the dilformer block containing multi-dilconv sparse attention (MDSA) and multi-dilconv feed-forward network (MDFN). The MDSA calculates the multi-scale query to generate accurate similarity map so that rich multi-scale information can be better utilized for the high-quality image reconstruction. In addition, we adopt ReLU to replace the original softmax to enforce sparsity in the Transformer for better feature aggregation. The MDFN is further established to better integrate the rain information of different scales in the feature transformation. Extensive experiments on the benchmarks show the favorable performance against state-of-the-art approaches.

1. Introduction

Single image deraining is a typical signal processing problem emerging in the last decade, whose aim is to recover the clean and rain-free background from its rain-degraded version. As rain streaks and clear image are unknown, it is a challenging ill-posed problem. Early studies [12, 15] usually solve this problem by performing a mathematical statistic to obtain a roughly generalized prior, but these priors are hard to align well with the real-world rain distribution, limiting their practical application.

With the success of the deep learning techniques, convolutional neural networks (CNNs)-based approaches [7, 11, 14, 24, 35, 38] have emerged for image deraining task and verified better restoration performance than those of traditional algorithms. These CNN-based methods have greatly advanced the progress thanks to the elaborately-designed network architectures and learning strategies [4]. However, these approaches still suffer from performance bottlenecks due to the intrinsic characteristics of the convolutional operations, *i.e*, local receptive fields and independence of input content, which limits the ability to eliminate long-range rain streaks.

Recently, Transformers [6,29,32,37], as the new network backbone, have achieved significant improvements boost over CNN models, because they can better model the non-local information for high-quality image reconstruction. Albeit these methods have achieved initial success, we observe that they fail to recover the fine spatial details of images and even invole implausible artifacts, especially in heavy rainy conditions highly degraded by intensive rain streaks. In fact, since the rain streak layer and rain-free background layer are highly interlaced, global and local representation learning are equally important for the challenging image deraining task, while the self-attention in Transformer does not manipulate the local invariance that CNNs do well. Thus, it is of great need to develop a hybrid network that combines the features by CNN and Transformer to obtain more robust deraining performance.

Towards this goal, we propose a dilated convolutional Transformer (DCT) for image deraining. Importantly, our approach incorporates dilated convolution operators with Transformer, thereby enlarging the receptive fields and producing contextually enriched feature representations. Specifically, the heart in DCT is the dilformer block, which covers two well-designed components. First, as not all the tokens from the queries are relevant to those in keys, using all similarity relations does not effectively facilitate the high-quality image reconstruction. Here, we develop a multi-dilconv sparse attention (MDSA) to select the most useful similarity values for global feature aggregation. On the other hand, existing Transformer-based deraining methods [32, 37] lack the utilization and exploitation of multi-scale rain information. In our model, a multi-dilconv feed-forward network (MDFN) is further established to better

*Corresponding author.

characterize the local rain streaks distribution. With above-mentioned designs, our proposed hybrid architecture can not only enrich the locality but also empower the capability of global feature exploitation, in order to facilitate rain removal.

The contributions of this paper are summarized as follows:

- We design an effective multi-dilconv sparse attention in Transformer, which is capable of generating more accurate representation for chasing high-quality deraining outputs.
- We design a novel multi-dilconv feed-forward network based on multi-scale fusion, where the rain information can be fully exploited to enrich inter-level feature transformation.
- Extensive experimental results on the commonly used benchmarks considerably demonstrate that our proposed DCT outperforms existing state-of-the-art deraining approaches.

2. Related Work

In this section, we will briefly introduce the recent related works for single image deraining and vision Transformer.

2.1. Single Image Deraining

Existing methods for single image deraining can be divided into two categories: prior-based methods and deep learning-based methods. Early deraining approaches generally develop different image priors to provide additional constraints. By correctly framing rain removal as an image decomposition challenge based on morphological component analysis, Kang et al. [12] construct a single picture deraining framework. Li et al. [15] propose a straightforward patch-based prior approach for modeling both the background and rain layers, achieving effective results.The deep learning-based deraining has shown excellent performance. Fu et al. [8] first introduce DetailNet that directly remove the rain layer by reducing the mapping range. Yang et al. [34] propose a multi-stage joint rain detection and estimate network and discuss the possible aspects as attribute and loss that effected on the deraining task. By utilizing convolutional and recurrent neural networks, RESCAN [14] proposes a way to make full use of contextual information for image rain removal. Zhang et al. [39] present a density-aware multi-stream densely connected CNN algorithm, DID-MDN, for estimating rain density on rainstreaks. In [24], Progressive Resnet Network (PReNet) carries out the recursive compute to effectively produce the derained images progressively. Jiang et al. [11] introduce a multi-scale progressive fusion network (MSPFN) for single image rain streak removal. RCDNet [26] utilizes a convolution dictionary to depict rain features and streamlined the network using proximal gradient descent technology. Wang et al. [27] develop a multi-decoding structure allows for optimal deraining features to be generated in each feature space by imposing individual supervision. Zou et al. [42] propose a novel data-free compression framework for deraining networks. Xiao et al. [32] present an efficient and effective transformer-based architecture for image deraining, which can capture long-range and complicated rainy artifacts.

2.2. Vision Transformer

Google introduces the Transformer [25], which exhibits exceptional performance in natural language processing (NLP). Numerous endeavors have been made to investigate the usage of the Transformer model in computer vision tasks, owing to the Transformer model's triumph in NLP. Recent studies shows that Transformers have achieved great success in high-level vision tasks such as image classification [1, 16, 18], segmentation [33, 41] and object detection [5, 9, 10]. Due to its outstanding performance, Transformer models have been studied for low-level vision tasks [2, 3, 20, 30]. Liang et al. [17] propose a strong baseline model SwinIR for image restoration based on the Swin [18] Transformer. Wang et al. [29] introduce a high-performing Transformer-based network called Uformer for image restoration. Lee et al. [13] present a attention mechanism for image restoration, named k-NN Image Transformer (KiT). Zamir et al. [37] suggest an effective Transformer for image restoration that can handle large images while modeling global connections. Transformers are better than CNNs at identifying long-range connections within data thanks to their global self-attention dependencies. Therefore, developing a hybrid network that combines CNN and Transformer features is crucial for achieving more robust deraining performance.

3. Proposed Method

In this section, our DCT architecture is presented first in this section (see Fig. 1). After describing the basic elements of the proposed dilformer block, we discuss its major components.

3.1. Overall Framework

Given an input 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 utilize a standard 3×3 convolutional layer as the projection of input and output. Our network architecture consists of a series of stacked $N_{i\in[1,2,3,4]}$ encoder/decoder dilformer units, which allows us to extract rich information for rain distribution that varies throughout space. It is necessary to concatenate the encoder features with the decoder features

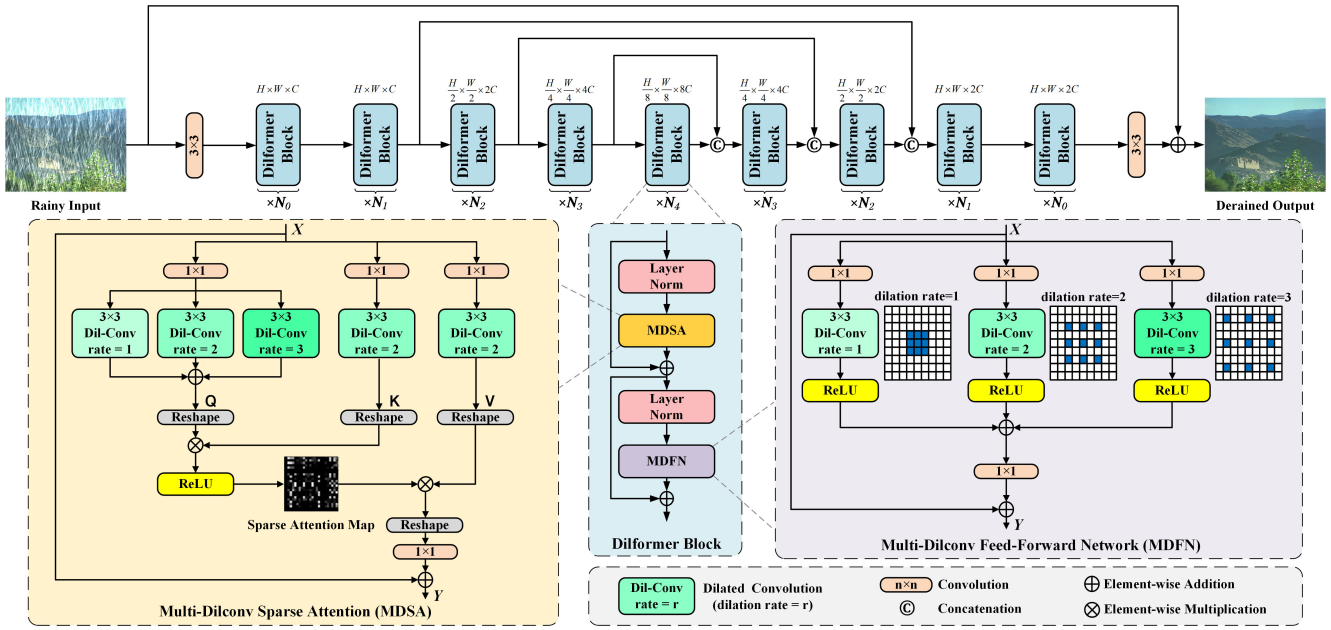

Figure 1. The overall architecture of the proposed dilated convolutional Transformer (DCT) for image deraining.

by means of skip connections in order to facilitate the recovery process. Instead of directly predicting a deraining image I_{derain}, the proposed model predicts a residual image I_{res} to which the degraded input image is added to obtain: $I_{derain} = I_{rain} + I_{res}$.

In each dilformer block, given the input features at the (l-1)-th block $\mathbf{X}_{l-1}$, the encoding procedures of dilformer block can be reformulated as

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

$$\mathbf{X}_l = \mathbf{X}'_l + \text{MDFN}\left(\text{LN}\left(\mathbf{X}'_l\right)\right), \tag{2}$$

where $\mathbf{X}'_l$ and $\mathbf{X}_l$ represent the outputs from the multi-dilconv sparse attention (MDSA) and multi-dilconv feed-forward network (MDFN). A layer normalization is referred to as an LN.

For simplicity, we leverage the pixel-wise loss by the $L1$ function to impose supervision on the learning process. We optimize our DCT end-to-end with the following objective:

$$\mathcal{L}_{pixel} = \left\| I_{derain} - I_{gt} \right\|_1 \tag{3}$$

where I_{derain} and I_{gt} denote the output derained image and the ground-truth image, respectively.

3.2. Multi-Dilconv Sparse Attention

Rain streaks can be automatically identified and removed using contextual information from an input image for image deraining [34]. Our method utilizes a contextualized dilated architecture rather than relying on depth-wise convolution [37] to aggregate context information at multiple scales for the purpose of learning rain features. This dilated convolution increases the receptive field without sacrificing resolution by weighting pixels with a step size of the dilated factor [36]:

$$g[i] = \sum_{l=1}^{L} f[i + r \cdot l] h[l], \tag{4}$$

where $f[i]$ is the input signal, $g[i]$ is the output signal , $h[l]$ denotes the filter of length L, and r corresponds to the dilation rate we use to sample $f[i]$. In standard convolution, $r = 1$.

The MDSA utilizes linear projection by first applying 1×1 convolution. In the next step, MDSA employs three dilated convolutions to compute the multi-scale query. Despite sharing kernel weights, these convolutions have different dilation rates: 1, 2, 3. The scale-independent similarities are then added together using a weighted tally. The three expanded pathways we use all use dilated convolutions with the same-sized kernel 3×3. Within the framework of self-attention, this design employs the usage of multi-scale data for calculating query and key similarity between any pair of spatial locations.

Furthermore, we also note that the softmax normalization in Transformer will keep all the similarities of the tokens from the query and key. However, not all the tokens from the query are relevant to those in key. Using the softmax normalization to generate self-attention would affect the following feature aggregation. As the ReLU is an effective activation function that can remove negative features while keep the positive ones, we use the ReLU to keep

Table 1. Comparison of quantitative results on five benchmark datasets. Bold and underline indicate the best and second-best results.

Datasets	Test100 [40]		Rain100H [34]		Rain100L [34]		Test2800 [8]		Test1200 [39]		Average	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DerainNet [7]	22.77	0.810	14.92	0.592	27.03	0.884	24.31	0.861	23.38	0.835	22.48	0.796
SEMI [31]	22.35	0.788	16.56	0.486	25.03	0.842	24.43	0.782	26.05	0.822	22.88	0.744
DIDMDN [39]	22.56	0.818	17.35	0.524	25.23	0.741	28.13	0.867	29.95	0.901	24.64	0.770
UMRL [35]	24.41	0.829	26.01	0.832	29.18	0.923	29.97	0.905	30.55	0.910	28.02	0.880
RESCAN [14]	25.00	0.835	26.36	0.786	29.80	0.881	31.29	0.904	30.51	0.882	28.59	0.858
PReNet [24]	24.81	0.851	26.77	0.858	32.44	0.950	31.75	0.916	31.36	0.911	29.43	0.897
MSPFN [11]	27.50	0.876	28.66	0.860	32.40	0.933	32.82	0.930	32.39	0.916	30.75	0.903
MPRNet [38]	30.27	0.897	30.41	0.890	36.40	0.965	33.64	0.938	32.91	0.916	32.73	0.921
DGUNet [23]	30.32	0.899	30.66	0.891	37.42	0.969	33.68	0.938	33.23	0.920	33.06	0.923
KiT [13]	30.26	0.904	30.47	0.897	36.65	0.969	33.85	**0.941**	32.81	0.918	32.81	0.926
Uformer-B [29]	29.90	0.906	30.31	**0.900**	36.86	0.972	33.53	0.939	29.45	0.903	32.01	0.924
IDT [32]	29.69	0.905	29.95	0.898	37.01	0.971	33.38	0.937	31.38	0.908	32.28	0.924
Ours	**30.91**	**0.912**	**30.74**	0.892	**38.19**	**0.974**	**33.89**	**0.941**	**33.57**	**0.926**	**33.46**	**0.929**

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

Methods	Rainy Image	MSPFN [11]	MPRNet [38]	DGUNet [23]	Ufomer-B [29]	IDT [32]	Ours
NIQE / BRISQUE	5.961 / 34.147	4.947 / 33.027	4.821 / 32.116	4.419 / 27.654	4.537 / 28.619	4.227 / 26.237	**4.103 / 24.795**

the most useful attentions for feature aggregation, automatically ensuring the sparse property of the attention weight.

Formally, we respectively apply a reshaping function to the query Q, key K, and V and obtain $\widehat{\mathbf{Q}} \in \mathbb{R}^{HW \times C}$, $\widehat{\mathbf{K}} \in \mathbb{R}^{HW \times C}$, and $\widehat{\mathbf{V}} \in \mathbb{R}^{HW \times C}$. Finally, we compute the sparse attention $\mathrm{A} \in \mathbb{R}^{C \times C}$ by:

$$\text{SparseAttention} = \mathrm{ReLU}\left(\frac{\widehat{\mathbf{Q}}^{\top}\widehat{\mathbf{K}}}{\alpha}\right)\widehat{\mathbf{V}}, \tag{5}$$

where α is a learnable parameter.

3.3. Multi-Dilconv Feed-Forward Network

Transformer allows the data to expand/reduce the dimension of the token and perform non-linear transformations on each token through the feed-forward network. Here, we propose a MDFN to enhance locality and increase the size of the receptive field in order to retrieve more contextual information. In fact, rich multi-scale representation has been fully demonstrated its effectiveness [11] in better removing rain. Similar to MDSA, the representations of the three convolution routes with varying dilation factors and receptive fields are aggregated to provide the final output features. MDFN also has the ability to apply random rates of dilation during the process, automatically expanding the receptive fields of the network without adding new modules, which is important for removing rainy effects of different appearances.

4. Experiments

In this section, we perform both quantitative and qualitative evaluations to validate the effectiveness of our proposed DCT on commonly used benchmark datasets. Here, we compare our method with 12 state-of-the-art image deraining approaches, including DerainNet [7], SEMI [31], DIDMDN [39], UMRL [35], RESCAN [14], PReNet [24], MSPFN [11], MPRNet [38], DGUNet [23], KiT [13], Uformer-B [29], and IDT [32].

4.1. Datasets and Metrics

Following [11, 37], we conduct extensive experiments on the Rain13K training dataset which contains 13, 700 clean/rain image pairs. For testing, five synthetic benchmarks (Test100 [40], Rain100H [34], Rain100L [34], Test2800 [8], and Test1200 [39]) are considered for evaluation. Note that, we calculate the PSNR and SSIM [28] scores using the Y channel in the YCbCr color space as quantitative comparisons. Besides, real-world datasets also considered to further evaluate the generalization performance. For the rainy images without clean labels, two no-reference indicators, NIQE [22] and BRISQUE [21], are employed for evaluating performance.

4.2. Implementation Details

The experiments are performed on PyTorch with 4 NVIDIA GTX 3090 GPUs. In our model, $\{N_0, N_l, N_2, N_3, N_4\}$ are set to $\{2, 4, 6, 6, 8\}$, and the number of attention heads for five dilformer blocks of the

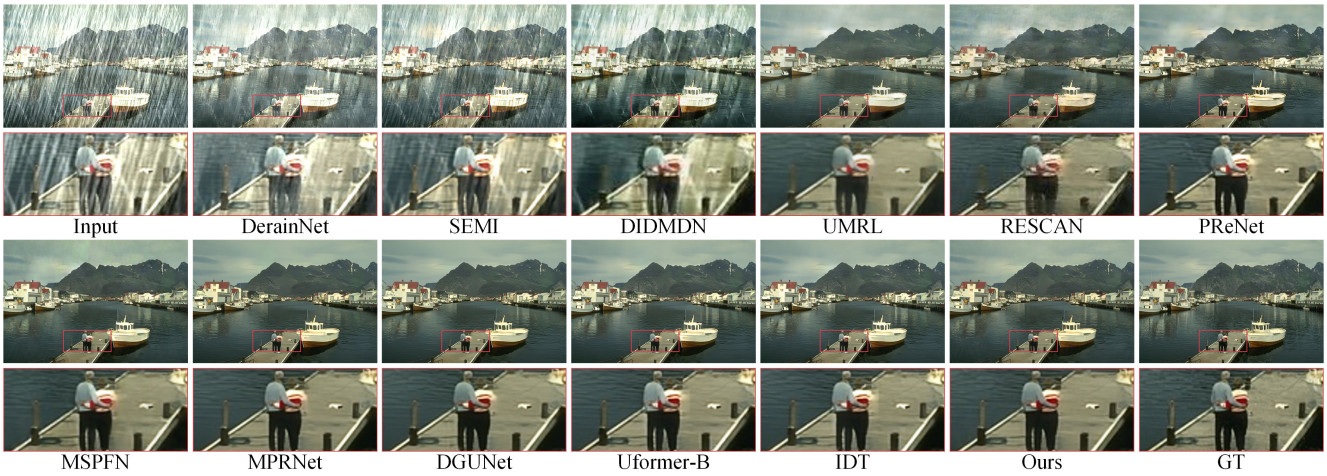

Figure 2. Visual quality comparison of deraining images obtained by different methods on the Rain100H benchmark dataset.

Figure 3. Visual quality comparison of deraining images obtained by different methods on the Test100 benchmark dataset.

Figure 4. Visual quality comparison of deraining images obtained by different methods on the Test1200 benchmark dataset.

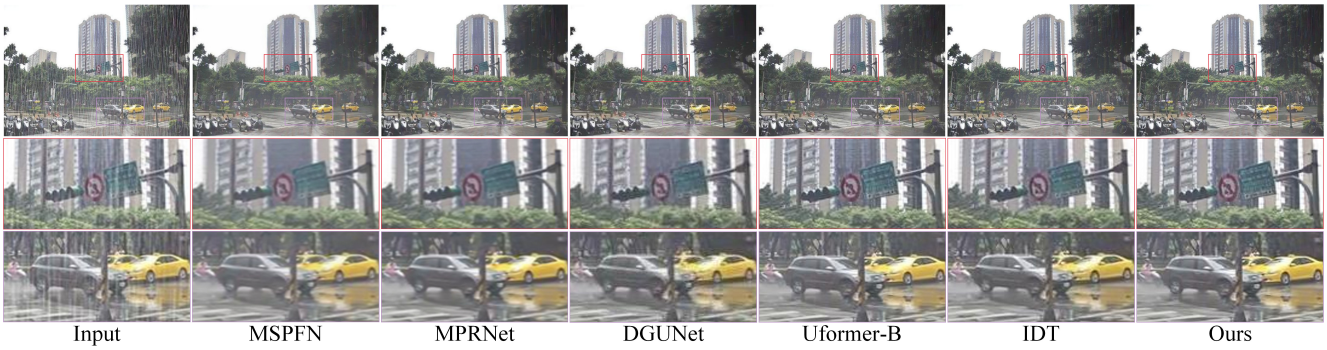

Figure 5. Visual quality comparison of deraining images obtained by different methods on real-world rainy images.

same level is set to $\{1, 1, 2, 4, 8\}$. The initial channel C is 48 and the expand ratio is set to 2. We use AdamW optimizer with batch size of 16 and patch size of 128 for total 300K iterations. The initial learning rate is fixed as 1×10^{-4} for 92K iterations, and then reduced to 1×10^{-6} for 208K iters with the cosine annealing [19].

4.3. Comparison with State-of-the-arts

Synthetic Datasets. Table 1 shows the quantitative evaluation results on the commonly used synthetic benchmarks. For fair comparisons, since UFormer-B and IDT are not trained on the Rain13K, we retrain the models using the default settings provided by the authors. For other methods, we evaluate them with their online codes. Obviously, our method gets the highest values both in PSNR and SSIM except for the SSIM of Rain100H, which can surely reflect the excellent performance and robustness of our designed DCT. As shown in Fig. 4 and Fig. 2, Test1200 and Rain100H images are provided for visual comparison. Fig. 3 shows the visual results on the Test100 dataset. Benefiting from the interaction of local and non-local information, our method can remove rain streaks while retaining more accurate details and credible textures in the background image.

Real-world Datasets. To further demonstrate the generalization of DCT, we compare it with other competing approaches on real-world dataset. As recored in Table 2, our net gets the lower NIQE and BRISQUE values, which means high-quality deraining results with clearer content

Figure 6. Ablation qualitative comparison for different variants of DCT. The models (a-e) are consistent with the settings in Table 3.

Table 3. Ablation study analysis on the Rain100H benchmark dataset.

Models	(a)	(b)	(c)	(d)	(e)
Depth-wise Conv	✓				
Dilated Conv		✓	✓	✓	✓
Single-scale Query	✓	✓			
Multi-scale Query			✓	✓	✓
r = 2,2,2			✓		
r = 1,2,3				✓	✓
ReLU	✓	✓	✓	✓	
Softmax					✓
PSNR	29.57	29.79	30.05	**30.74**	30.21
SSIM	0.877	0.879	0.884	**0.892**	0.886

and better perceptual quality. Through the visual comparison in Fig. 5, our method can achieve better generalization performance and high-quality restoration in detail preservation and rain removal.

4.4. Ablation Studies

To demonstrate the superiority of our framework, we conduct studies on different variants of DCT. We mainly consider the following variants: (1) depth-wise or dilated convolution; (2) single-scale or multi-scale query; (3) same or different dilation rates; (4) ReLU or Softmax. The quantitative results on the Rain100H are listed in Table 3. We observe that our model (d) performs better than the other possible configurations, which shows that each design strategy we consider brings their own gains to the final performance of DCT. As shown in Fig. 6, our method (d) can generate a clearer recovery result.

5. Conclusion

In this paper, we have presented an effective dilated convolutional Transformer (DCT) for image deraining. We build the dilformer block with combinations of multi-dilconv sparse attention and multi-dilconv feed-forward network, and show that it can boost high-quality restoration performance significantly. Extensive experimental results show that the proposed DCT achieves favorable deraining performance against state-of-the-art methods.

Acknowledgements. This work was supported by Liaoning Provincial Applied Basic Research Project under Grant 2022JH2/101300247.

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021. 2

[2] Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, and Donglai Wei. Learning to generate realistic noisy images via pixel-level noise-aware adversarial training. *Advances in Neural Information Processing Systems*, 34:3259–3270, 2021. 2

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 2

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2017–2026, 2022. 1

[5] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2988–2997, 2021. 2

[6] 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. *arXiv preprint arXiv:2010.11929*, 2020. 1

[7] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network archi-

tecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017. 1, 4

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3855–3863, 2017. 2, 4

[9] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5321–5330, 2022. 2

[10] Chenhang He, Ruihuang Li, Shuai Li, and Lei Zhang. Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8417–8427, 2022. 2

[11] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020. 1, 2, 4

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

[13] Hunsang Lee, Hyesong Choi, Kwanghoon Sohn, and Dongbo Min. Knn local attention for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2139–2149, 2022. 2, 4

[14] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 254–269, 2018. 1, 2, 4

[15] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2736–2744, 2016. 1, 2

[16] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 2

[17] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2

[18] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[20] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–466, 2022. 2

[21] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 4

[22] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 4

[23] Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17399–17410, 2022. 4

[24] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2019. 1, 2, 4

[25] 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

[26] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3103–3112, 2020. 2

[27] Yinglong Wang, Chao Ma, and Bing Zeng. Multi-decoding deraining network and quasi-sparsity based training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13375–13384, 2021. 2

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

[29] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 1, 2, 4

[30] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 2

[31] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3877–3886, 2019. 4

[32] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 4

[33] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 2

[34] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1357–1366, 2017. 2, 3, 4

[35] Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8405–8414, 2019. 1, 4

[36] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[37] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 1, 2, 3, 4

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 1, 4

[39] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 695–704, 2018. 2, 4

[40] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*, 30(11):3943–3956, 2019. 4

[41] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[42] Weiqi Zou, Yang Wang, Xueyang Fu, and Yang Cao. Dreaming to prune image deraining networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6023–6032, 2022. 2