Multi-scale fusion of gated neighborhood attention transformers for single image deraining

Liu, Y.; Xiao, G.; Lew, M.S.K.; Wu, S.

ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446762

MULTI-SCALE FUSION OF GATED NEIGHBORHOOD ATTENTION TRANSFORMERS FOR SINGLE IMAGE DERAINING

Yijin Liu[1], *Guoqiang Xiao*[1], *Michael S. Lew*[2], *Song Wu*[1,*]

[1]College of Computer and Information Science, Southwest University, China;
[2]LIACS, Leiden University, Netherlands

ABSTRACT

Since the diverse geometric appearances and densities of rain streaks, local-global information is equally essential for single image deraining. Balancing local-global information becomes a challenge. Thus, a Multi-Scale Fusion of Gated Neighborhood Attention Transformers (MSF-GNAT) for single image deraining is proposed in this paper. Firstly, a Gated Neighborhood Attention Transformer (GNAT) block is designed to achieve complex condition deraining by complementarily fusing global and local information. Secondly, a Multi-Scale Fusion (MSF) block is developed to fuse multi-scale information for richer representation. Moreover, a Simplified Gated Feed-forward Network (SGFN) is proposed to ensure that the convolutional results primarily attend to valid pixels. Evaluation experiments on synthetic and real-world datasets demonstrate the superiority of our MSF-GNAT over state-of-the-art methods. The source code is available at https://github.com/SWU-CS-MediaLab/MSF-GNAT.

Index Terms— Deraining, Transformer, Multi-Scale Fusion, Neighborhood Attention

1. INTRODUCTION

Image deraining task refers to removing raindrops from an image to using image processing techniques, resulting in a cleaner and clearer output image. Although image deraining is a typical low-level visual task, it is crucial for many outdoor high-level visual tasks such as object tracking, video surveillance, and pedestrian detection. These advanced visual tasks require recognizing and understanding image data in clear and visible image scenes.

Due to the success of deep learning in advanced computer vision tasks, CNN-based methods were proposed [1]. However, the inherent operations limitations, such as the restricted local receptive field, hindered the ability to handle long-range dependencies. Consequently, the Transformer is applied in the area of image deraining [2]. The global interaction capability of the Transformer has achieved promising performance. However, the self-attention mechanism in Transformer has introduced new challenges, including increased complexity and a lack of local knowledge modeling. However, rich local-global information representations are both indispensable for deraining, and few methods can achieve the balance. Moreover, multi-scale feature learning has achieved good results in single image deraining[3], but these methods fail to explore the correlation between the original damaged image and cross-scale information.

To alleviate the above limitations, an efficient multi-scale fusion architecture based on the designed gated neighborhood attention transformer (MSF-GNAT) is proposed for single-image deraining. In MSF-GNAT, the designed Multi-Scale Fusion (MSF) block is employed to merge multi-scale features asymmetrically, aiming to explore a more comprehensive perceptual field that effectively aids in the restoration of image structure and details. Additionally, a Gated Neighborhood Attention Transformer (GNAT) block is developed to extract image features under uneven rainfall density. Specifically, we combine complementary Neighborhood Attention (NA) [4] and Dilated Neighborhood Attention (DINA) [5] to capture fine-grained local features for image detail restoration, while capturing long-range dependencies for better overall visual effects in image restoration. Furthermore, to optimize the learned model to focus on the informative regions and differentiate between image background and raindrops, a Simplified Gated Feed-forward Network (SGFN) is designed in the GNAT to enhance information flow and achieve high-quality outputs. The main contributions of our MSF-GNAT are summarised as follows:

- The designed Multi-Scale Fusion (MSF) block employs a cascaded mechanism to integrate multi-scale information progressively, and it can effectively leverage the cross-scale correlation of rain streaks.
- The developed Gated Neighborhood Attention Transformer (GNAT) architecture can capture global information and improve locality. The Simplified Gated Feed-forward Network (SGFN) in GNAT can optimize the information flow to enhance the feature representation capability of the learned model.
- Comprehensive performance evaluation experiments conducted on various benchmark datasets demonstrated the superiority of our MSF-GNAT.

*Corresponding author

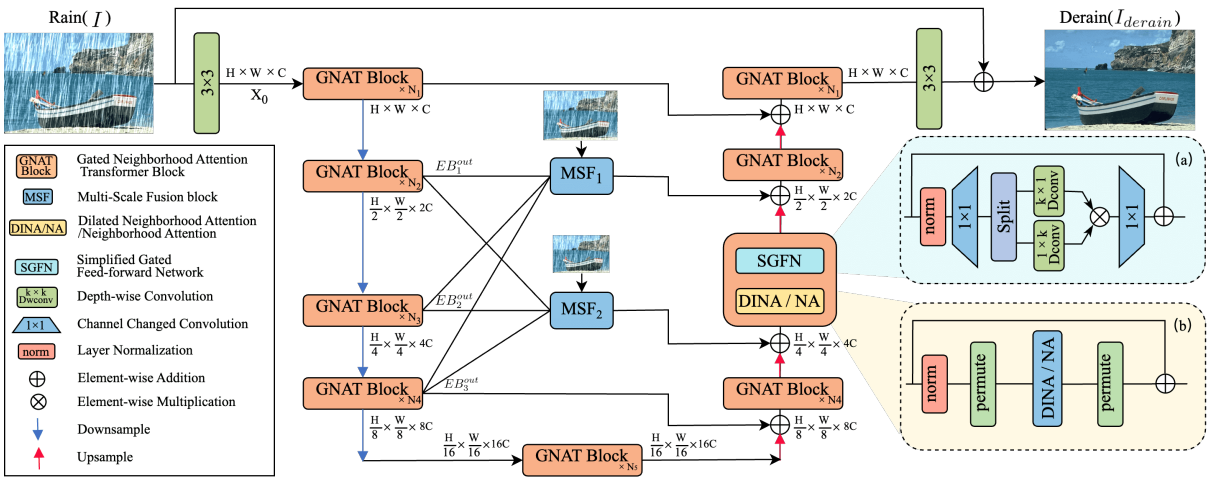

Fig. 1: The demonstration of the detailed framework of our proposed MSF-GNAT. GNAT block and MSF are two critical modules in MSF-GNAT. The core modules in the GNAT architecture are SGFN and DINA/NA. (a) SGFN is designed to control information flow. (b) DINA/NA is introduced to capture long-range dependencies while modeling locality.

2. PROPOSED METHOD

As shown in Fig. 1, a U-shaped encoder-decoder framework is adopted in our Multi-Scale Fusion of Gated Neighborhood Attention Transformers (MSF-GNAT). Specifically, given a rain image $I \in R^{H\times W\times 3}$, shallow features $X_0 \in R^{H\times W\times C}$ are extracted through 3×3 convolution and fed into the encoder and decoder. In the backbone network, an amount of N_i Gated Neighborhood Attention Transformer (GNAT) blocks are stacked at each stage to extract rich rain streak features with spatial variations. Pixel-unshuffle and pixel-shuffle operations are employed for downsampling and upsampling, respectively. To better fuse multi-scale information, the encoded image features are fused by passing through the Multi-Scale Fusion (MSF) block before being fed into the decoder. $I_{derain} = I + \mathcal{F}(I)$ gives the final reconstructed structure, where $\mathcal{F}(\cdot)$ represents the entire network.

2.1. Gated Neighborhood Attention Transformer Block

Fig. 1 (a) and (b) present the details of the GNAT block. GNAT block consists of three sub-modules. The first module is the Neighborhood Attention (NA) [4] module, which computes attention only among neighboring pixels to enhance local modeling capability. The second module is the Dilated Neighborhood Attention (DINA) [5] module, which expands the receptive field and captures long-range dependencies to learn global structures. The third module is the Simplified Gated Feed-forward Network (SGFN), which controls information flow and strengthens the feature representations. Layer normalization is applied before the attention operations, and residual connections are used to enhance the feature learning. Formally, given the input features X_l^{m-1} at the N_i^l block, the encoding process of the GNAT block can be defined as:

$$\begin{aligned} \hat{x}_l^m &= x_l^{m-1} + NA(LN(x_l^{m-1})) \\ x_l^m &= \hat{x}_l^m + SGFN(LN(\hat{x}_l^m)) \\ \hat{x}_l^{m+1} &= x_l^m + DINA(LN(x_l^m)) \\ x_l^{m+1} &= \hat{x}_l^{m+1} + SGFN(LN(\hat{x}_l^{m+1})) \end{aligned} \tag{1}$$

Where x_l^{m-1} and x_l^{m+1} denote the input and output features of GNAT block, respectively.

Neighborhood Attention. The rain effect includes rain streaks. Rain streaks cause reflections with neighboring images, resulting in the loss of image information caused by specular highlights. To recover the lost information, exploring more significant local information is necessary. Therefore, we introduce the NA mechanism as the local attention to capture local information and better restore and complete the image information lost due to rain streaks. The NA mechanism can be formulated as follows:

$$NA_k(i) = softmax(\frac{A_i^k}{\sqrt{d}})V_i^k \tag{2}$$

Where $\sqrt{d}$ is the scaling parameter. $NA_k(i)$ represents the neighborhood attention of the i-th token with a neighborhood size of k. A_i^k and V_i^k represent the weights and value of the i-th token, respectively.

Dilated Neighborhood Attention. The rain effect in the real world also includes the accumulation of rain bands, which can result in a mist-like visual in the distance. Thus, more rich features must be explored in a smaller field of view to help restore this information. To address this, DINA is introduced as the sparse global attention to capture global information and obtain long-range dependencies. The DINA mechanism can be formulated as follows:

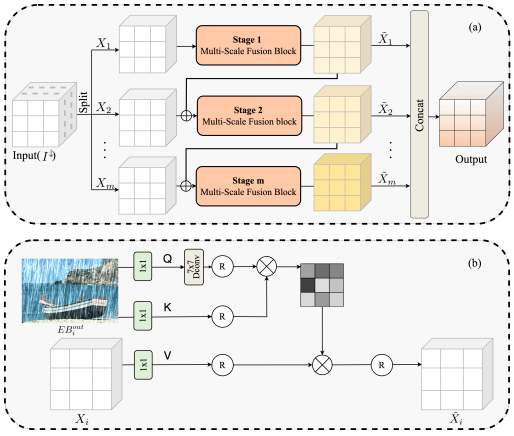

Fig. 2: (a) shows the architecture of the MSF block. The Stage i Multi-Scale Fusion block details are illustrated in (b).

$$DINA_k^{\delta}(i) = softmax(\frac{A_i^{(k,\delta)}}{\sqrt{d}})V_i^{(k,\delta)} \tag{3}$$

Where $DINA_k^{\delta}(i)$ denotes the δ-dilated neighborhood attention of the i-th token with a neighborhood size of k.

Simplified Gated Feed-forward Network. As previously mentioned [6, 7], the standard Feed-forward Network lacks the capability of control over information flow. Thus, the designed SGFN aims to optimize the information flow. For the given input $\hat{X}_l$, a convolution with shape 1×1 is applied to expand the channel dimension (expansion factor=2.66). The tensor is then split along the channel dimension into $\hat{X}_l^1$ and $\hat{X}_l^2$, which fed into two parallel branches. Each branch performs a $k \times 1$ and $1 \times k$ depth-wise convolution to extract features and better capture spatial information. A simplified gated mechanism is achieved through element-wise multiplication, and finally, a 1×1 convolution is used to reduce the channel dimension back to the original size. Thus, the computation of SGFN can be represented as:

$$\begin{aligned}
[\hat{X}_l^1, \hat{X}_l^2] &= Split(Conv_{1\times1}(\hat{X}_l)) \\
\hat{X}_l^h &= DWConv_{k\times1}(\hat{X}_l^1) \\
\hat{X}_l^w &= DWConv_{1\times k}(\hat{X}_l^2) \\
X_l &= Conv_{1\times1}(\hat{X}_l^h \odot \hat{X}_l^w)
\end{aligned} \tag{4}$$

Where $Conv_{1\times1}$ represents 1×1 convolution , $DWConv_{1\times k}$ and $DWConv_{k\times1}$ denote $1 \times k$ and $k \times 1$ depth-wise convolution, respectively, and $\odot$ indicates element-wise multiplication.

2.2. Multi-Scale Fusion Block

Previous research has indicated that incorporating multi-scale information can enhance the effectiveness of single image rain removal [8, 9]. However, more than simple fusion operations are required to utilize complementary information from different scales effectively. Therefore, inspired by this insight [10], we designed the MSF block. In the MSF block, the upsampled or downsampled features from the encoder block (EB) simultaneously at different scales block and the downsampled rainy image are used as the inputs; thus, learning and leveraging the complementary information between these inputs can effectively guide the image restoration process. The specific encoding process is as follows:

$$\begin{aligned}
MSF_1^{out} &= MSF(EB_1^{out}, (EB_2^{out})^{\uparrow}, (EB_3^{out})^{\uparrow}, I^{\downarrow}) \\
MSF_2^{out} &= MSF((EB_1^{out})^{\downarrow}, EB_2^{out}, (EB_3^{out})^{\uparrow}, I^{\downarrow})
\end{aligned} \tag{5}$$

MSF_n^{out} denotes the n-th output of the MSF block, and EB_n^{out} represents the output of EB at the n-th level. Operations of up-sampling ($\uparrow$) and down-sampling ($\downarrow$) are applied such that the features from different scales can be concatenated.

As shown in Fig. 2(a), the channels are first split into m stages for fusion (m=3), progressively refining the feature representation. The fusion details are illustrated in Fig. 2 (b), and finally, concatenation is performed along the channel dimension. The specific computation of MSF block is as follows:

$$\begin{aligned}
X_i &= X_i + \tilde{X}_{i-1} \\
\tilde{X}_i &= Attn(EB_i^{out}W^Q, EB_i^{out}W^K, X_iW^V) \\
MSF_n^{out} &= Concat[\tilde{X}_i]_{i=1:m}
\end{aligned} \tag{6}$$

2.3. Loss Function

The commonly used loss function in network training is L_1 loss. However, L_1 loss emphasizes low-frequency information in images and weakens high-frequency features. Therefore, Fast Fourier Transform (FFT) loss is incorporated to preserve high-frequency details [8]. Thus, the comprehensive loss function of our model can be represented as follows:

$$\mathcal{L} = \|I_{derain} - I_{gt}\|_1 + \lambda\|\mathcal{F}(I_{derain}) - \mathcal{F}(I_{gt})\|_1 \tag{7}$$

where $\mathcal{F}$ denotes the FFT, I_{gt} represents the target image, I_{derain} denotes the output image from the deraining network, $\|\cdot\|_1$ denotes the L_1-norm and λ is set to 0.1.

3. EXPERIMENTS

3.1. Measures and Datasets

To evaluate the effectiveness of our MSF-GNAT, we conducted experiments on three synthetic datasets (Rain200L [11], Rain200H [11], DID-Data [12]) and one real-world dataset (SPA-data [13]). We compared our method against the state-of-the-art techniques, including MSPFN [9], IDT [14], RCDNet [15], SPDNet [16], DualGCN [17], Restormer [6], and DRSformer [18]. We employed two commonly used metrics, Peak Signal-to-Noise Ratio (PSNR) [19] and Structure Similarity (SSIM) [20], for quantitative comparison. Furthermore, to validate the generalization of our MSF-GNAT, the performance on the Internet-Data dataset [13] is also tested, which does not contain ground-truth images.

3.2. Implementation Details

MSF-GNAT is implemented in the PyTorch framework using the AdamW optimizer. During the training process, we utilized an NVIDIA GeForce RTX 4090 GPU with a patch size of 128, iterating a total of 300K times.

Table 1: Average SSIM and PSNR comparison results on Rain200L, Rain200H, DID-Data and SPA-Data. The best results are **bolded**, and the second best results are underlined.

Datasets	Venue/Year	Rain200L		Rain200H		DID-Data		SPA-Data	
Metrics		SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
MSPFN	CVPR2020	38.58	0.9827	29.36	0.9034	33.72	0.9550	43.43	0.9843
RCDNet	CVPR2020	39.17	0.9885	30.24	0.9048	34.08	0.9532	43.36	0.9831
SPDNet	ICCV2021	40.50	0.9875	31.28	0.9207	34.57	0.9560	43.20	0.9871
DualGCN	AAAI2021	40.73	0.9886	31.15	0.9125	34.37	0.9620	44.18	0.9902
IDT	TPAMI2022	40.74	0.9884	32.10	0.9344	34.89	0.9623	47.35	0.9930
Restormer	CVPR2022	40.99	0.9890	32.00	0.9329	35.29	0.9641	47.98	0.9921
DRSformer	CVPR2023	41.23	0.9894	32.18	0.9330	35.38	0.9647	**48.53**	0.9924
MSF-GNAT	**Ours**	**41.57**	**0.9901**	**32.44**	**0.9362**	**35.59**	**0.9667**	48.33	**0.9935**

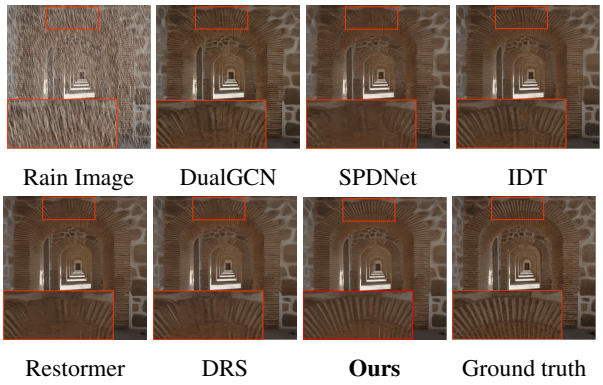

Fig. 3: Visual quality comparison on DID-Data dataset.

3.3. Results on Synthetic Datasets

Table 1 presents the quantitative results of different methods on various synthetic datasets, including Rain200L, Rain200H, and DID-Data. It can be observed that our proposed MSF-GNAT achieves the best performance in terms of SSIM and PSNR metrics. Fig. 3 shows that while the comparative methods remove a significant portion of rain streaks, they also result in the loss of some original texture information. In contrast, our restored images exhibit more apparent edge details and image textures.

3.4. Results on Real-world Datasets

To validate the effectiveness of our method in practical applications, we compare the performance of all competing approaches in the last column of Table 1 on the SPA-Data dataset. The results show that our method can compete with previously popular rain removal methods. From the visual quality comparison in Fig. 4 and Fig. 5, it can be observed that our process removes a significant portion of rain streaks. In contrast, other state-of-the-art methods still retain some noticeable rain artifacts.

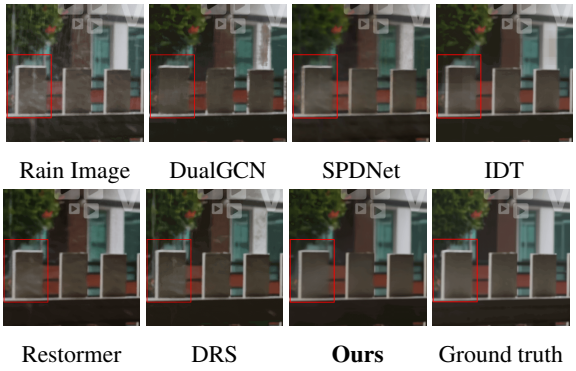

Fig. 4: Visual quality comparison on SPA-Data dataset.

Fig. 5: Visual quality comparison on Internet-Data dataset.

3.5. Ablation Study

To further investigate the effectiveness of different components in MSF-GNAT, we conducted comprehensive ablation experiments on the Rain200L dataset with consistent parameter settings. We verified the usefulness of the proposed MSF block and SGFN, as well as the introduced DINA/NA. Experimental results can be shown in Table 2.

Table 2: Comparison results with/without the proposed MSF block and SGFN and the introduced DINA/NA modules.

Modules	Baseline	Baseline+MSF	Baseline+MSF+SGFN	Ours
PSNR	35.74	37.39	39.42	41.57
SSIM	0.9719	0.9789	0.9855	0.9901

4. CONCLUSION

This paper proposes an MSF-GNAT framework for single image deraining. The developed MSF block adopts a cascade operation to fuse multi-scale information in different channels, and it can explore and use the complementary features between different scales. The designed GNAT block uses alternate NA and DINA to obtain long-distance dependencies for local information modeling. Meanwhile, the designed SGFN controls information flow and suppresses the spread of useless features. Performance comparison to SOAT shows the effectiveness of our MSF-GNAT in both quantitative and qualitative evaluations, and the ablation study shows the advantages of each designed module in MSF-GNAT.

5. REFERENCES

[1] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2007–2016.

[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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12299–12310.

[3] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8343–8352.

[4] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi, "Neighborhood attention transformer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6185–6194.

[5] Ali Hassani and Humphrey Shi, "Dilated neighborhood attention transformer," *ArXiv*, vol. abs/2209.15001, 2022.

[6] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5718–5729.

[7] Liangyu Chen, Xiaojie Chu, X. Zhang, and Jian Sun, "Simple baselines for image restoration," *European Conference on Computer Vision*, 2022.

[8] Sung-Jin Cho, Seoyoun Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4621–4630.

[9] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8343–8352.

[10] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14420–14430.

[11] Wenhan Yang, Robby T. Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu, "Joint rain detection and removal from a single image with contextualized deep networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 6, pp. 1377–1393, 2020.

[12] He Zhang and Vishal M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 695–704.

[13] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson W.H. Lau, "Spatial attentive single-image deraining with a high quality real rain dataset," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12262–12271.

[14] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha, "Image de-raining transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2022.

[15] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng, "A model-driven deep neural network for single image rain removal," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3100–3109.

[16] Qiaosi Yi, Juncheng Li, Qinyan Dai, Faming Fang, Guixu Zhang, and Tieyong Zeng, "Structure-preserving deraining with residue channel prior guidance," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4218–4227.

[17] Xueyang Fu, Qi Qi, Zheng-Jun Zha, Yurui Zhu, and Xinghao Ding, "Rain streak removal via dual graph convolutional network," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 1352–1360.

[18] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan, "Learning a sparse transformer network for effective image deraining," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5896–5905.

[19] Quan Huynh-Thu and Mohammed Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.

[20] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.