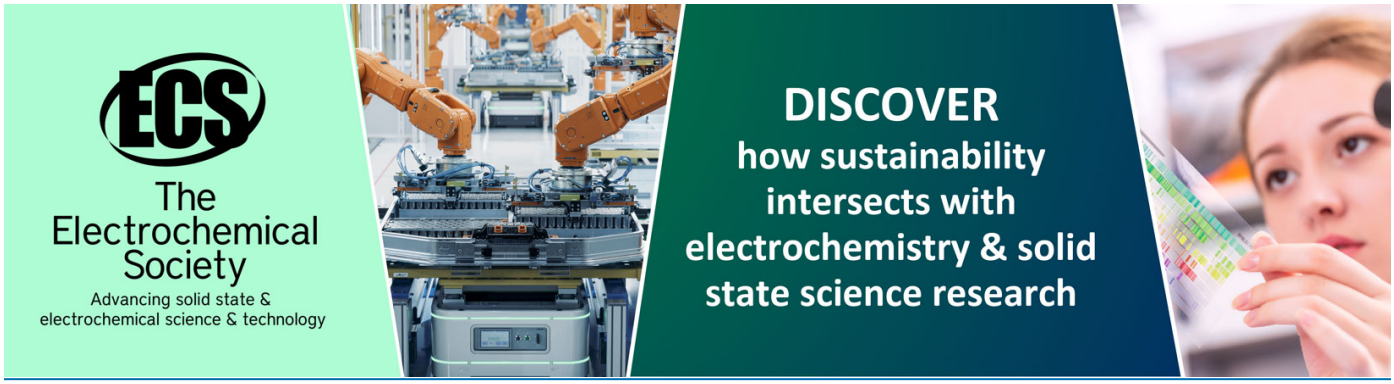
Single Image Rain Removal Algorithm Based on U-Net and Vision Transformer



Single Image Rain Removal Algorithm Based on U-Net and Vision Transformer

Hua Cui[1], Zhiqiang Yang[1] and Yanzi Shi[2,3]

[1]School of Information Engineering, Chang' an University, Xi' an, China

[2]Corresponding author, School of Automobile, Chang' an University, Xi' an, China

[3]yzshi@chd.edu.cn

Abstract. Rain will significantly lower image quality, which will undoubtedly have an impact on how well outdoor computer vision systems operate, such as autonomous driving. This paper proposes a two-branch deep neural network consisting of an attention guided U-Net and Vision Transformer, which could capture intra-field details and cross-patch relationships to obtain good global and local deraining effects. In order to ensure both branches pose positive effects on the target derain task, we specifically design a patch boot image module to achieve complemented feature fusion, which adaptively selects informative intra-field regions guided by the patch-wise importance map. Moreover, Wasserstein distance between the prediction and reference image is applied as the objective function to pursue better image quality measurement. In comparison to existing algorithms, the suggested method may more effectively remove rain and restore the background details, according to qualitative and quantitative results on the public datasets Rain200L and Rain200H. On the aforementioned datasets, the peaks of the structural similarity (SSIM) and signal-to-noise ratio (PSNR) were 32.55/0.9476 and 26.12/0.8826, respectively.

1. Introduction

Due to the substantial quality deterioration of images taken in rainy circumstances, several computer vision algorithms, including object identification, image segmentation, and depth estimation [1], which are essential components of autonomous navigation and surveillance systems, perform poorly. In order to improve the dependability of these vision systems, it is essential to remove the effects of rainy weather from the photographs. The difficulty of removing rain from a single picture is made even more difficult by the fact that the location and timing of the single image are static and the distribution of rain markings in rainy photographs is usually variable. Therefore, a single image draining offers a great deal of practical value. Early methods usually impose various prior knowledge based on the statistical properties of rain marks and clean images, which limits the rain removal performance due to the difficulty of simulating complex and changeable rainy weather scenes with prior knowledge [2].

Recently, many deep learning-based approaches have achieved satisfactory performance [3], but CNN's convolutional layers cannot directly capture the correlation of local pixels, leading to constrained receptive fields. Stacking convolution kernels is how most of the known deraining models increase the receptive field. However, the receptive field obtained in this way is still limited, and the global information of the rain image is not effectively used, resulting in the loss of some details and structural information of the image after deraining.

Transformers have been applied to images deraining with good performance [4]. These techniques, however, are unable to accurately model the local features of the image or recover its local information.

We propose a dual-branch deep neural network made up of the Attention guided U-Net and Vision Transformer to achieve good global and local deraining effects. The former network employs spatial attention to choose valuable local information from low-level characteristics under the supervision of high-level features, while the later network uses a transformer to gather global information. The main contributions of this study are summarized as follows:

(1) To obtain good global and local deraining effects, we propose a two-branch deep neural network consisting of an attention guided U-Net and Vision transformer, which could capture intra-field details and cross-patch relationships.

(2) To fuse features derived from two branches, we design a patch boot image module to adaptively select informative intra-field regions for the target derain task.

(3) Wasserstein distance between the prediction and reference image is applied as the objective function to pursue better image quality measurement.

2. Related Work

2.1. The basic structure of U-Net

The segmentation performance of medical images is significantly improved by the U-Net network suggested by Renneberger *et al.* [5]. U-Net's fundamental structure is divided into two parts. The first part is called encoder equipped with several same block network, which consists of two consecutive 3×3 convolutions, followed by a ReLU function and one max pooling layer. The second part is the decoder consisting of reverse blocks with the same number as that of the encoder, where each block first uses 2×2 up-convolution to up-sample the feature map, then the feature map corresponding to the encoder is cropped and connected to the up-sampled feature map, followed by two 3×3 convolution and ReLU function. To get the feature map down to the necessary number of channels and create a segmented image, one additional 1×1 convolution is employed in the final stage. The U-Net network has a nearly symmetrical form and a U-like appearance.

2.2. Transformer

The early applications of Transformer were for natural language processing (NLP), and they were successful. The encoder and decoder subnetworks make up this network. In the encoder stage, the words in the sentence are first converted into word vectors, then feature map of global attention is obtained through self-attention module and finally output of the encoder is obtained through feedforward network. The output of the corresponding encoder and the output of the preceding decoder are included in the input of a decoder. Since the parallel input lacks the positional relationship of words, Transformer uses positional encoding to preserve the positional relationship, and the output of the decoder is the probability distribution of the corresponding position. As a result of the aforementioned noteworthy achievement, Transformer is being used in the field of computer vision by an increasing number of researchers. The Vision Transformer model was proposed by Dosovitskiy *et al*. [6]. Transformer is being used for the first time to categorise images. In order to adjust to the encoder input, Vision Transformer separates the image into non-overlapping image patches. A new patch is included to predict the final label at the output of the Transformer encoder in a manner similar to BERT's [class] token.

3. Proposed Method

We first explain the overall architecture of our suggested method in this section, and then introduce the specifically designed Attention guided U-Net, Vison Transformer and Patch boot image module.

3.1. overall architecture

Figure 1 depicts the overall architecture. Given the rainy image $Input \in R^{h \times w \times c}$, where $h \times w$ represents the spatial resolution, $input$ is sent to the Attention guided U-Net to obtain local information.

After $Input$ is divided into several patches, global information a and region coefficient b are obtained through Vision Transformer. Finally, we use the Patch boot image module to fuse the local information from U-Net with global information a and region coefficient b from Vision Transformer to obtain $Output \in R^{h\times w\times c}$. By minimizing the following loss function, our model is trained:

$$\mathcal{L} = f(Output, truth), \tag{1}$$

where f stands for the Wasserstein distance and $truth$ stands for the ground-truth image.

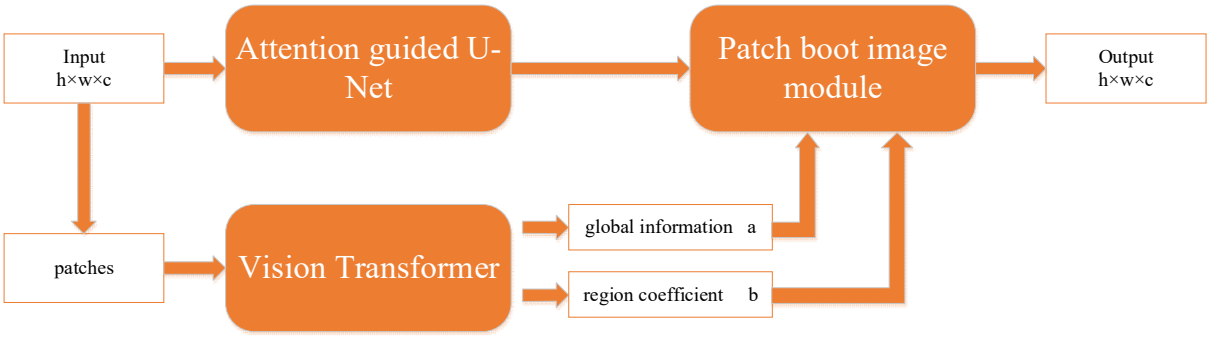


Figure 1. overall architecture.

3.2. Attention guided U-Net

In general, low-level features have more specific data, while high-level features have more semantic data. For the single image deraining work, both features are essential. Most U-Net-based methods directly connect the features of different levels, which do not consider the characteristics of different levels of features, and, generate some irrelevant features affecting the performance of the network structure. Therefore, we propose spatial attention guidance module (SAG). **Figure 2**'s right side demonstrates it. Pass the high-level f_h and low-level f_l features to the spatial attention guidance module (SAG) in order to produce the cascaded feature $\hat{f}_l$. By conducting an up-convolution operation on the high-level feature f_h, SAG first creates the attention map A for the low-level features:

$$A = SP(f_h'), \tag{2}$$

where f_h' stands for the features f_h after up-convolution and $SP(\cdot)$stands for spatial attention block. From f_l, the attention maps are used to extract the important information.

$$f_l' = A \otimes f_l, \tag{3}$$

where f_l' stands for the refined low-level feature with attention map A and $\otimes$ is pointwise multiplication. For the encoder, we suggest stacking high-level features created by up-convolution with refined low-level features. The 2D convolutional layer is then used to obtain the output feature $\hat{f}_l$.

$$\hat{f}_l = Conv\big(Concat(f_h', f_l')\big), \tag{4}$$

where $Conv(\cdot)$ stands for the convolution layer with a 1×1 kernel size and $Concat(\cdot)$ for the concatenation operation. Be note that a SAG module relates to each Attention directed U-Net layer.

3.3. Vision Transformer

Figure 3 depicts the Vision Transformer's structural layout. picture $Input \in R^{h\times w\times c}$ is divided into many flattened uniformly non-overlapping patches $x_p \in R^{N\times(p^2\cdot c)}$, where $p \times p$ stands for the

dimension of each patch and $N = \left[\frac{h \times w}{p^2}\right]$ is the length of the picture sequence, to prepare the input data for Vision Transformer. Through a patch encoder, we project patches onto the K-dimensional embedding space.

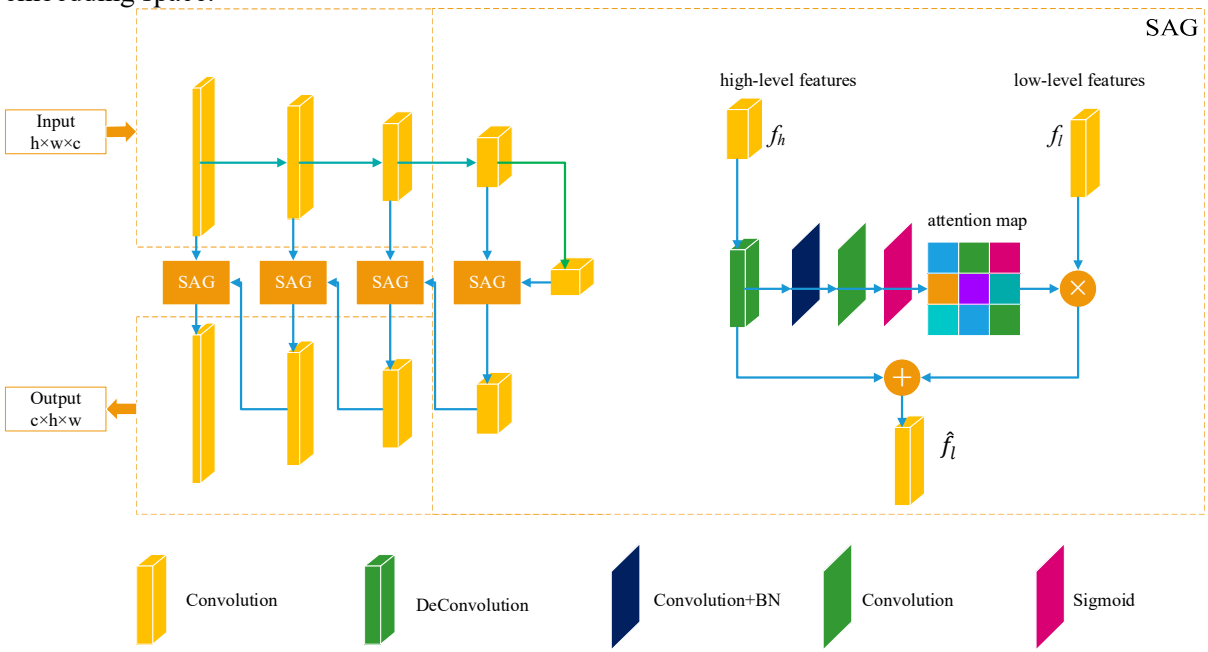


Figure 2. Attention guided U-Net.

We first train a 1D position embedding $position \in R^{N \times K}$, which is then added to the patch embedding to preserve the position information. This preserves the spatial information of each patch. $t_0 = [x_p^1 I; x_p^{21} I; \cdots; x_p^N I] + position$, where $I \in R^{(p^2 \cdot c) \times K}$ denotes the projected patch embedding.

To learn global information, we employ a stack of Transformer blocks made up of multi-head self-attention (MSA) and multi-layer perceptron (MLP). To scale the embedded patches, the MSA layer consists of L parallel self-attention heads: $t_i' = MSA\big(Norm(t_{i-1})\big) + t_{i-1}, i = 1 \cdots L$. By using the formulas $t_i = MLP\big(Norm(t_i')\big) + t_i', i = 1 \cdots L$,where $Norm(\,)$denotes layer normalisation and $t_i \in R^{N \times d}$ denotes the encoded semantic representation in d-dimensional space, the MLP module learns global information. We model global information $a \in R^{64 \times h \times w}$ in addition to encoding characteristics by reshaping t_L and using a 1×1 convolution technique.

In order to describe the significance of pixels in each region, we additionally define the region coefficient b. The region coefficient b 's function is to give the Patch boot image module a supervisory signal to help it select significant regions. The region coefficient $b \in R^{1 \times N}$is also obtained by reshaping t_L and applying a 1 × 1 convolution operation.

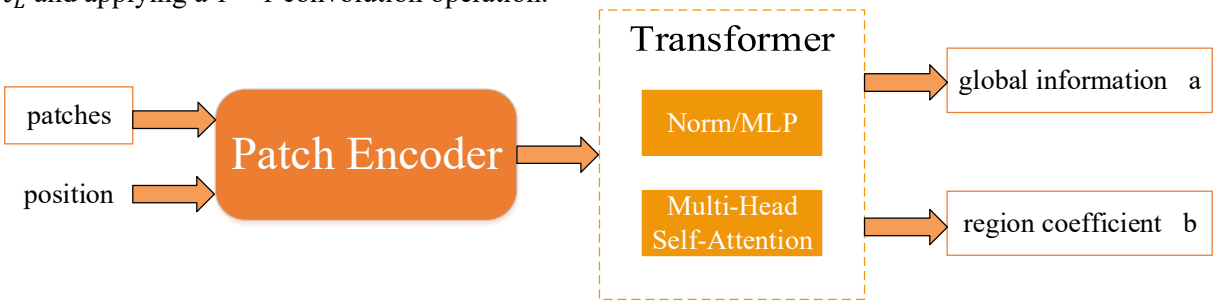


Figure 3. Vision Transformer.

3.4. Patch boot image module

This module is shown in **Figure 4**. x is the output from Attention guided U-Net, which represents local information. The global information a and region coefficient b are outputs of Vision Transformer. First, multiply x by region coefficient b, and then connect the multiplication result with the global information a in the channel dimension to obtain the fusion feature. Finally, the $output$ is obtained through a series of convolution operations and sigmoid activation functions.

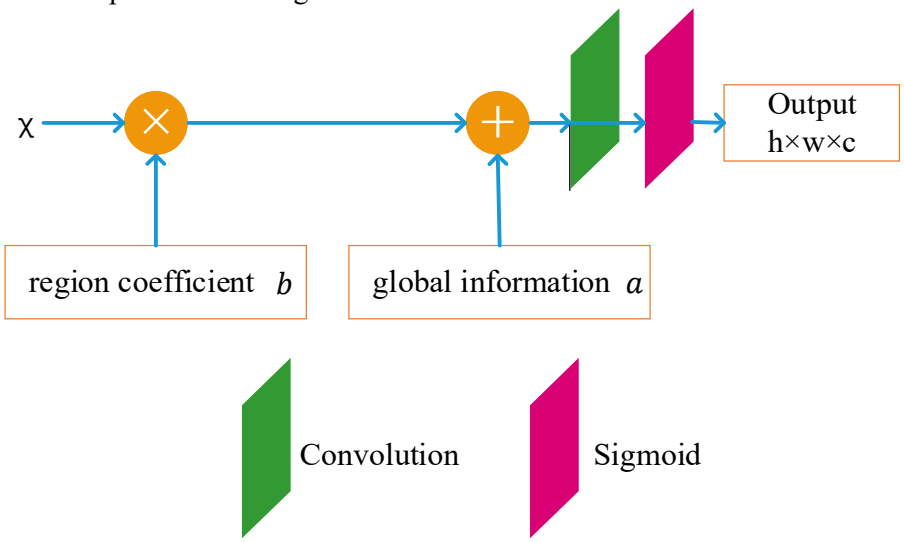


Figure 4. Patch boot image module

4. Experiments and Analysis

We introduce the datasets, evaluation measures, comparison techniques, implementation specifics, and experimental findings in this part.

4.1. Datasets

We apply the deraining experiments on the Rain200L and Rain200H public benchmarks. 1800 simulated images for training and 200 synthetic images for testing are included in Rain200L and Rain200H, respectively.

4.2. Evaluation metrics

Peak Signal-to-Noise Ratio ($PSNR$) and Structure Similarity Index ($SSIM$) are employed as the assessment metrics for the aforementioned benchmarks. In accordance with the earlier deraining approach, we calculate the $PSNR$ and $SSIM$ in the Y channel of the $YCbCr$ space.

4.3. Comparison methods

We contrast the suggested method with DSC [7], GMM [8], and DDN [9], the findings of which are referenced in [10].

4.4. Implementation details

On a single NVIDIA RTX 3060 GPU utilized for training, the experiments are carried out using the Pytorch framework. We use the Adam optimizer with a batch size of 4 and a learning rate of 0.0075. Moreover, if the validation accuracy does not increase for 10 successive epochs, a scheduler for learning rates is utilised to lower the learning rate in half. 200 epochs were used to train the network.

4.5. Results

4.5.1. Quantitative result. The quantitative assessment of the Rain200L and Rain200H is presented in **Table 1**. The proposed method performs better than previous comparing methods, as evidenced by increased PSNR and SSIM. On the Rain200L dataset, our technique performs better than DSC by 5.39 dB on PSNR and 0.0813 on SSIM. It also performs better than GMM by 3.89 dB on PSNR and 0.0824 on SSIM. Our technique outperforms DSC on the Rain200H dataset by 11.39 dB on PSNR and 0.5011 on SSIM, while outperforming GMM by 11.62 dB on PSNR and 0.4662 on SSIM. Such a big improvement demonstrates that our suggested strategy significantly enhances images deraining performance.

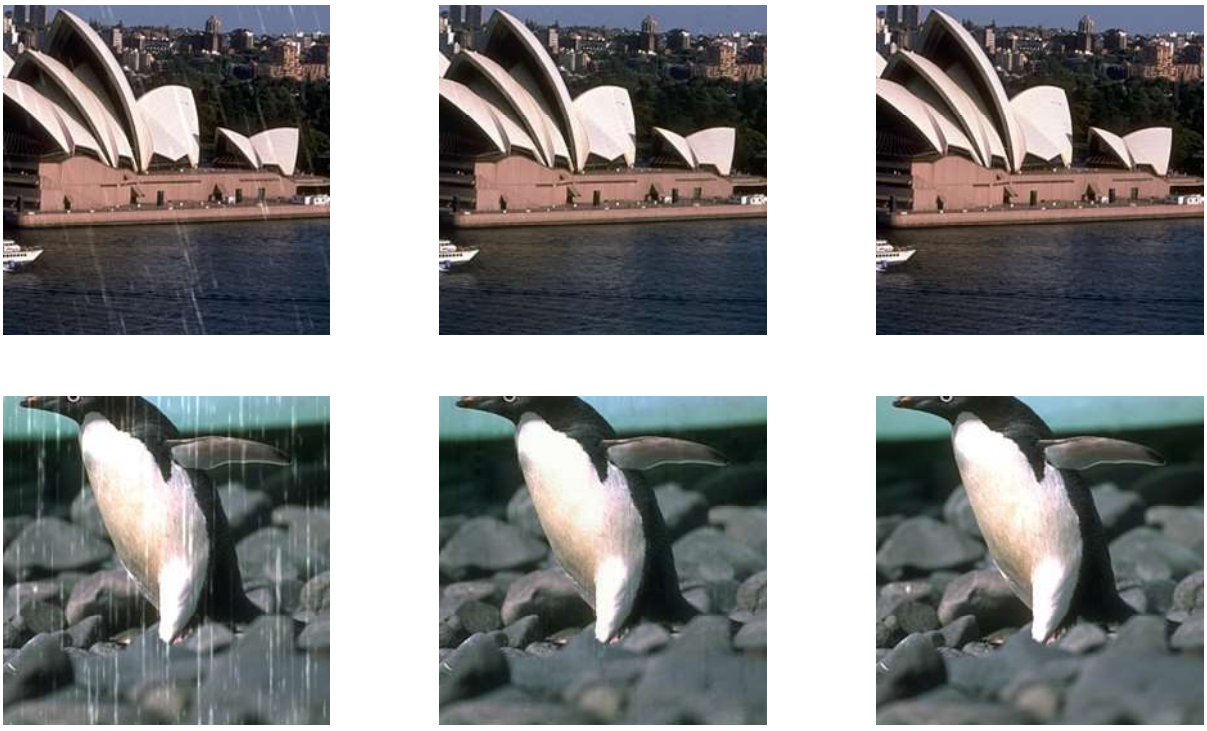


Figure 5. Visual results on Rain200L

4.5.2. Qualitative result. **Figures 5 and Figures 6** display the visual deraining outcomes for Rain200L and Rain200H, respectively. It is evident that our approach can produce effective outcomes for removing rain that are almost in line with reality.

Table 1. Quantitative comparison between the Rain200L and Rain200H. The top two outcomes are denoted by **bold** and underline, respectively.

Datasets		Rain200L		Rain200H	
Metrics		PSNR	SSIM	PSNR	SSIM
Method	DSC	27.16	0.8663	14.73	0.3815
	GMM	28.66	0.8652	14.50	0.4164
	DDN	**34.68**	**0.9671**	26.05	0.8056
	Ours	32.55	0.9476	**26.12**	**0.8826**

Figure 6. Visual results on Rain200H

5. Concluding Remarks

This paper proposes a dual-branch deep neural network composed of attention guided U-Net and Vision Transformer. To ensure that both branches have a positive impact on the target deraining task, we specifically design a patch boot image module for complemented feature fusion, which adaptively selects informative intra-field regions guided by the patch-wise importance map. Furthermore, the

Wasserstein distance between the predicted image and the reference image is used as an objective function in pursuit of a better image quality measurement. Experimental results and analysis demonstrate the effectiveness of our algorithm.

6. References

[1] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16* (pp. 213-229). Springer International Publishing.

[2] Kang, L. W., Lin, C. W., & Fu, Y. H. (2011). Automatic single-image-based rain streaks removal via image decomposition. *IEEE transactions on image processing*, *21*(4), 1742-1755.

[3] Chen, X., Pan, J., Jiang, K., Li, Y., Huang, Y., Kong, C., ... & Fan, Z. (2022). Unpaired deep image deraining using dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2017-2026).

[4] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., ... & Gao, W. (2021). Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12299-12310).

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (pp. 234-241). Springer International Publishing.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[7] Luo, Y., Xu, Y., & Ji, H. (2015). Removing rain from a single image via discriminative sparse coding. In *Proceedings of the IEEE international conference on computer vision* (pp. 3397-3405).

[8] Li, Y., Tan, R. T., Guo, X., Lu, J., & Brown, M. S. (2016). Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2736-2744).

[9] Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., & Paisley, J. (2017). Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3855-3863).

[10] Chen, X., Li, H., Li, M., & Pan, J. (2023). Learning A Sparse Transformer Network for Effective Image Deraining. *arXiv preprint arXiv:2303.11950*.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (Grant Nos. 12171234) and the Fundamental Research Funds for the Central Universities, CHD (300102223104).