IEICE TRANS. INF. & SYST., VOL.E106–D, NO.7 JULY 2023

LETTER

A Fusion Deraining Network Based on Swin Transformer and Convolutional Neural Network

Junhao TANG[†], ***Nonmember and*** **Guorui FENG**[†a)], ***Member***

SUMMARY Single image deraining is an ill-posed problem which also has been a long-standing issue. In past few years, convolutional neural network (CNN) methods almost dominated the computer vision and achieved considerable success in image deraining. Recently the Swin Transformer-based model also showed impressive performance, even surpassed the CNN-based methods and became the state-of-the-art on high-level vision tasks. Therefore, we attempt to introduce Swin Transformer to deraining tasks. In this paper, we propose a deraining model with two sub-networks. The first sub-network includes two branches. Rain Recognition Network is a Unet with the Swin Transformer layer, which works as preliminarily restoring the background especially for the location where rain streaks appear. Detail Complement Network can extract the background detail beneath the rain streak. The second sub-network which called Refine-Unet utilizes the output of the previous one to further restore the image. Through experiments, our network achieves improvements on single image deraining compared with the previous Transformer research.

key words: *Swin Transformer, convolutional neural network, multi-scale dilated convolution, single image deraining*

1. Introduction

With the development of industry cameras and monitoring video, the requirements of deraining in images are constantly growing. Removing rain streaks and restoring background content are needed to obtain better background scene for downstream computer vision tasks and human perception. Recent methods [1], [2] were mostly based on convolutional neural network (CNNs), which achieved impressive results. Moreover, some probes attempt to utilize Transformer [3], [4] to accomplish the rain removal task.

For single image deraining, the traditional method and the deep-learning based method are the two main categories. Recently, the deraining method based on Transformer is emerging and has achieved great performance improvement on benchmark public datasets. *Traditional methods:* Some researches concentrate on physical peculiarity of the rain streaks and background scenes with a well-designed optimization model. Typical methods include information of frequency [7], modeling streaks via Gaussian Mixture Model (GMM) [8], normalizing rain layers with Discriminative Sparse Coding (DSC) [9]. *Deep-Learning-based methods:* Some researches utilize residual information to accomplish deraining tasks [10]. Authors in [11] introduce attention mechanism to better recovery the image. In [12], authors enhance the image quality via progressive structural boosting constraints. In [21], a well-designed autocoder is made for embedding supervision. The model can get great performance on severely-degraded datasets.

Transformer-based methods: Authors in [3] focus on building a single encoder-decoder. They propose a transformer-based encoder with intra-patch transformer (Intra-PT) blocks, which focuses on attention inside the main patches to remove the degradation. Although the Transformer model overcomes the defects of CNNs (*i.e.*, limited receptive field), the complexity of computation increases quadratically with spatial resolution. In that case, Restormer [4] is proposed, which is computationally efficient. To reduce computational complexity, authors introduce multi-Dconv head transposed attention (MDTA) module and gated-Dconv feed-forward network (GDFN). These two networks can model global context and perform controlled feature transformation respectively. Nevertheless, the baseline which based on Transformer can achieve better performance when cooperating with a recovery mechanism based on CNNs. Therefore, we attempt to fuse Transformer and CNNs to complete deraining tasks. Recently, authors in [5] present Swin Transformer, which achieves great performance improvement on image classification. Moreover, in more computer vision tasks including image segmentation [5] and inpainting [6], methods that adopt Swin Transformer have surpassed those based on CNNs.

In this paper, we propose a new approach to perform single image deraining via a fusion network based on Swin Transformer and convolutional neural network. The major contributions of our work are summarized as follows. 1) We utilize a Unet with Swin Transformer for single image deraining. The mentioned network works as detecting the location of rain streaks and outputting the preliminary rain-free image. 2) We design a multi-scale feature complement network to improve the performance on texture restoration. In order to reduce the loss of texture information, we extract features of different scales.

2. Proposed Method

The proposed network is based on Swin Transformer. There are two sub-networks in the overall framework as shown in Fig. 1. The first sub-network has two branches, including Rain Recognition Network and Detail Complement Network. The second sub-network is Refine-Unet.

Manuscript received February 9, 2023.
Manuscript revised April 3, 2023.
Manuscript publicized April 24, 2023.

[†]The authors are with School of Communication and Information Engineering, Shanghai University, China.
a) E-mail: grfeng@shu.edu.cn

DOI: 10.1587/transinf.2023EDL8009

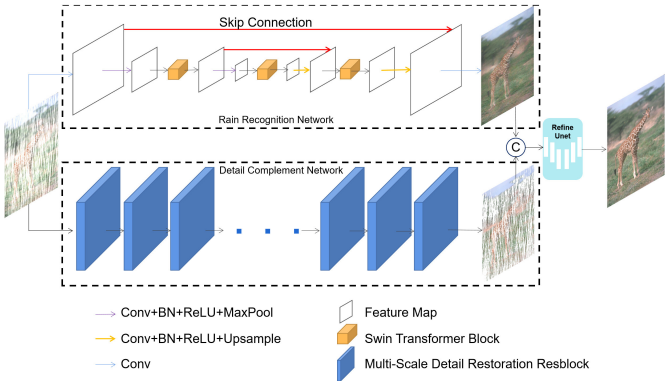

Fig. 1 The structure of proposed network. The network takes the rainy image as input and outputs the clean background. There are two main sub-networks corresponding to different functions. The first sub-network is two-branch architecture, including Rain Recognition Network and Detail Complement Network. Rain Recognition Network is a Unet embedded with Swin-Transformer. Detail Complement Network consists of 20 Multi-Scale Detail Restoration Resblocks in series. The second sub-network is Refine-Unet shown in Fig. 4.

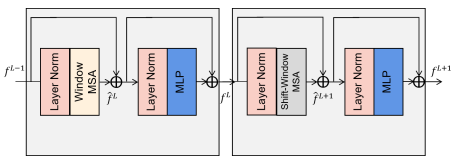

Fig. 2 The structure of two Swin Transformer Layers (STLs).

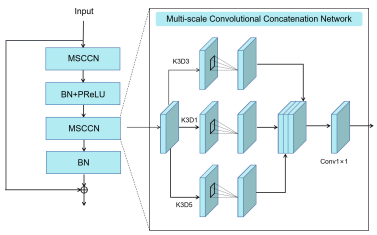

Fig. 3 The structure of Multi-Scale Detail Restoration Resblock (MSDRRB).

2.1 Rain Recognition Network

Although Transformer-based methods has excellent performance on specific vision tasks, there are two main problems: 1) Transformer is inferior when modeling the long sequence because the computational complexity will increase quadratically with the spatial resolution. 2) Transformer does poor in dealing with the tasks like instance segmentation [13]. However, Swin Transformer well solves the mentioned two problems. Due to the addition of shifted-window, the parameter of networks is decreased thus improving the performance on pixel-wise vision tasks.

To let the network focus on global feature, in Rain Recognition Network, we adopt Swin Transformer Block (STB) to replace the common-used convolution layer. As shown in Fig. 2, the number of Swin Transformer Layers (STLs) is multiples of two, each of which includes window multi-head self-attention (W-MSA) and shifted-window multi-head self-attention (SW-MSA). As described above, there exists matters when directly appling Transformer on vision tasks. Thus, the author proposed cyclic shift to reduce operation time while preserving the features of convolution. In our network, one STB includes eight Swin Transformer Layers. Processes of two STLs are respectively represented as:

$$\begin{aligned} \hat{f}^{L} &= W - MSA(LN(f^{L-1})) + f^{L-1}, \\ f^{L} &= MLP(LN(\hat{f}^{L})) + \hat{f}^{L}, \end{aligned} \tag{1}$$

$$\begin{aligned} \hat{f}^{L+1} &= SW - MSA(LN(f^{L})) + f^{L}, \\ f^{L+1} &= MLP(LN(\hat{f}^{L+1})) + \hat{f}^{L+1} \end{aligned} \tag{2}$$

where $LN(.)$ represents Layer Normalization. MLP is multi-layer perceptron in which two layers with Gaussian Error Linear Unit (GELU) are connected.

2.2 Detail Complement Network

Inspired by [14], we design our Detail Complement Network based on the multi-scale aggregated recurrent Resnet. To fully utilize multi-scale features of background beneath the rain streaks, we adopt 20 Multi-Scale Detail Restoration Resblock (MSDRRB) which are connected in series to form the complete Detail Complement Network. As shown in Fig. 3, to acquire larger receptive field in order to get more texture information, we adopt dilated convolutions with one, three and five dilation scales respectively. To recovery the dimension of feature maps to the input dimension, we add a K3D1 (Kernel 3 and Dilation 1) convolution [14] in the end of the sub-block. The output of Detail Complement Network can complement the rain-free image from Rain Recognition Network, providing richer texture information for further restoration.

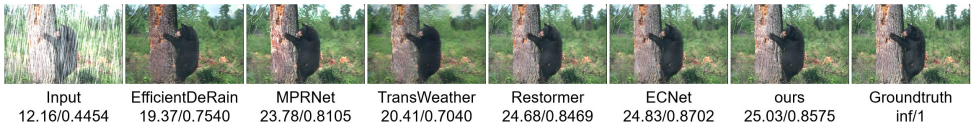

Fig. 5 The deraining performance comparisons on Rain200H.

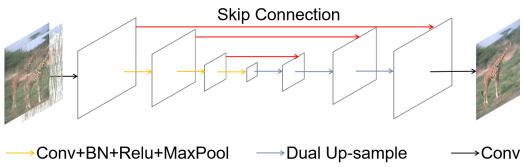

Fig. 4 The structure of Refine-Unet. We replace the common-used transpose convolution with dual Up-sample when generating the feature map.

2.3 Refine Unet

Normally, Up-sample in Unet is realized by transpose convolution. Inspired by [15], to reduce transmission loss during convolution, we use dual Up-sample module which consists of Bilinear and PixelShuffle [16]. The improvement brought by dual Up-sample will be illustrated in Chapter 3.

2.4 Loss Function

We train our network in an end-to-end manner with Mean Square Error (MSE) loss for image deraining:

$$Loss_1 = \|O_1 - B\|_2^2 \tag{3}$$

where O_1 denotes the result of Rain Recognition Network and B is the target rain-free image. $Loss_1$ focuses on the difference between the target image and the rain-free image output from Rain Recognition Network.

$$Loss_2 = \|O_1 + O_2 - B\|_2^2 \tag{4}$$

where O_2 denotes the result of Detail Complement Network. $Loss_2$ focuses on the difference between the target image and the combination of O_1 and O_2.

$$Loss_3 = \|O_3 - B\|_2^2 \tag{5}$$

where O_3 denotes the final output of the entire network. B is the target rain-free image. $Loss_3$ focuses on the difference between the target image and the output of the entire work.

$$Loss = \kappa Loss_1 + \mu Loss_2 + \lambda Loss_3 \tag{6}$$

where κ, μ and λ are tradeoff parameters. In our work, we set κ, μ as 0.1 and λ is set as 1.

3. Experimental Results

3.1 Experimental Setups

For all experiments, we activate the network by using NVIDIA GeForce RTX 3090 GPUs, adopting Adam optimizer [17] with the batch size of 10 and the patch size of 128×128. The learning rate is 1×10^{-4}. The total epoch is 100. In our experiment, Rain200L [18] and Rain200H [18] are used to validate proposed method, which contain light and heavy synthetic rain-streaks degradation respectively. We employ the common-used Peak Signal to Noise Ratio (PSNR) [19] and Structural Similarity (SSIM) [20] to evaluate the result on synthetic datasets.

Table 1 Ablation study on different settings of our method on Rain200H.

	M1	M2	M3	M4	M5	M6
STB	✗	✓	✓	✓	✓	✓
MSDRRB	20	20	20	17	21	22
d-Up	✗	✗	✓	✓	✓	✓
PSNR	28.54	30.75	30.96	29.98	30.90	30.88
SSIM	0.8752	0.8981	0.9018	0.8876	0.9015	0.8996

3.2 Ablation Study

To show the effectiveness of deraining on severely degraded rainy images, we take the performance on Rain200H as demonstration of the ablation study. We adopt PSNR and SSIM to quantitatively analyze the performance. Table 1 shows the influences brought by different components.

- STB indicates using Swin Transformer Block to replace the common-used convolution in Rain Recognition Network.
- MSDRRB indicates the number of Multi-Scale Detail Restoration Resblock (MSDRRB) connected in series.
- d-Up indicates using dual Up-sample to replace transpose convolution in Refine-Unet.

The result proves that STB greatly promotes SSIM for making better use of global information. In addition, MSDRRB greatly promotes PSNR when considering larger receptive field. Moreover, for the reason that small amount of texture information still exists in the preliminary output with not thoroughly removed rain streaks, the final Refine-Unet seems to be particularly important. The introduction of dual Up-sample can reduce the loss when generating the final output.

3.3 Experiments on Synthetic Data

We compare our method with several recently proposed methods: TransWeather [3], Restormer [4], MPRNet [1], ECNet [21], and EfficientDeRain [22]. All these methods share the same training and testing datasets.

Compared with other methods, we fully consider both

Table 2 Quantitative results evaluated with respect to average PSNR and SSIM.

Datasets	Rain200L	Rain200H
Metrics	PSNR SSIM	PSNR SSIM
MPRNet [1]	36.40/0.9634	29.51/0.8902
EfficientDeRain [22]	35.63/0.9776	28.17/0.8837
TransWeather [3]	35.03/0.9675	26.78/0.8557
Restormer [4]	38.51/0.9743	30.11/0.8992
ECNet+LL [21]	38.86/0.9865	30.15/0.9101
ours	38.92/0.9798	30.96/0.9018

global and multi-scale feature, so that our method obtains the better score on PSNR. Moreover, the result of a simple baseline with Transformer on deraining tasks tends to be slightly inadequate. Combination of Transformer and CNNs can better cope with images with severe degradation. It should be noticed that the baseline of ECNet can get excellent performance on SSIM. The idea, of which enables the encoder-decoder to predict rain layer which is as close as groundtruth rain layer, can better derive accurate rain-free background.

4. Conclusions

This paper proposes a deraining model with the combination of Swin Transformer and CNNs. The proposed network concludes two sub-networks. The first sub-network includes two branches. Rain Recognition Network is a Unet with the Swin Transformer layer, which focuses on global feature of rain images. Detail Complement Network works as extracting the background detail beneath the rain streak, which focuses on multi-scale information. Refine-Unet utilizes the output of previous two branch networks to further restore the image. Experiments show that our network can achieve performance improvement on single image deraining compared with the previous researches based on Transformer.

References

[1] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.14821–14831, 2021.

[2] H. Wang, Q. Xie, Q. Zhao, and D. Meng, "A model-driven deep neural network for single image rain removal," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3103–3112, 2020.

[3] J.M.J. Valanarasu, R. Yasarla, and V.M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.2353–2363, 2022.

[4] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5728–5739, 2022.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.10012–10022, 2021.

[6] J. Jiang, J. Zhu, M. Bilal, Y. Cui, N. Kumar, R. Dou, F. Su, and X. Xu, "Masked Swin Transformer Unet for Industrial Anomaly Detection," IEEE Transactions on Industrial Informatics, vol.19, no.2, pp.2200–2209, 2023.

[7] L.-W. Kang, C.-W. Lin, and Y.-H. Fu, "Automatic single-image-based rain streaks removal via image decomposition," IEEE Transactions on Image Processing, vol.21, pp.1742–1755, 2011.

[8] Y. Li, R.T. Tan, X. Guo, J. Lu, and M.S. Brown, "Rain streak removal using layer priors," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.2736–2744, 2016.

[9] Y. Luo, Y. Xu, and H. Ji, "Removing rain from a single image via discriminative sparse coding," Proceedings of the IEEE International Conference on Computer Vision, pp.3397–3405, 2015.

[10] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3855–3863, 2017.

[11] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp.8022–8031, 2019.

[12] L. Peng, A. Jiang, H. Wei, B. Liu, and M. Wang, "Ensemble single image deraining network via progressive structural boosting constraints," Signal Processing: Image Communication, vol.99, 116460, 2021.

[13] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.8759–8768, 2018.

[14] W. Yang, R.T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Joint rain detection and removal via iterative region dependent multi-task learning," CoRR, abs/1609.07769, 2, 1–12, 2016.

[15] C.-M. Fan, T.-J. Liu and K.-H. Liu, "SUNet: Swin Transformer UNet for Image Denoising," IEEE International Symposium on Circuits and Systems (ISCAS), pp.2333–2337, 2022. doi: 10.1109/ISCAS48785.2022.9937486

[16] W. Shi, J. Caballero, F. Huszar, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," Proceedings of the IEEE conference on CVPR, pp.1874–1883, 2016.

[17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations, 2014.

[18] W. Yang, R.T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1357–1366, 2017.

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

[20] Z. 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.

[21] Y. Li, Y. Monno, and M. Okutomi, "Single Image Deraining Network with Rain Embedding Consistency and Layered LSTM," IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022.

[22] Q. Guo, J. Sun, F. Juefei-Xu, L. Ma, X. Xie, W. Feng, Y. Liu, and J. Zhao, "Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining," Proceedings of the AAAI Conference on Artificial Intelligence, vol.35, no.2, pp.1487–1495, 2021.