SANDFORMER: CNN AND TRANSFORMER UNDER GATED FUSION FOR SAND DUST IMAGE RESTORATION

Jun Shi[1†], *Bingcai Wei*[2†], *Gang Zhou*[1*], *Liye Zhang*[2]

[1]School of Information Science and Engineering, Xinjiang University
[2]College of computer science and technology, Shandong University of Technology

ABSTRACT

Although Convolutional Neural Networks (CNN) have made good progress in image restoration, the intrinsic equivalence and locality of convolutions still constrain further improvements in image quality. Recent vision transformer and self-attention have achieved promising results on various computer vision tasks. However, directly utilizing Transformer for image restoration is a challenging task. In this paper, we introduce an effective hybrid architecture for sand image restoration tasks, which leverages local features from CNN and long-range dependencies captured by transformer to improve the results further. We propose an efficient hybrid structure for sand dust image restoration to solve the feature inconsistency issue between Transformer and CNN. The framework complements each representation by modulating features from the CNN-based and Transformer-based branches rather than simply adding or concatenating features. Experiments demonstrate that SandFormer achieves significant performance improvements in synthetic and real dust scenes compared to previous sand image restoration methods.

Index Terms— Gate fusion, Sand image restoration, Transformer branch, CNN branch

1. INTRODUCTION

Sandstorms are one of the most common dynamic weather phenomena and can significantly reduce the visibility and contrast of captured images. The existing sand dust degraded image restoration methods are mainly based on the atmospheric light scattering model to improve the classic haze removal algorithm directly. Due to some assumptions and prior knowledge constraints, the processing results of sand dust images will have color cast and blur. In recent years, the use of deep learning methods to enhance haze images has achieved great success. Inspired by this, some scholars began to apply deep learning methods to sand dust image restoration.

[1] proposed a convolutional neural network sand dust image enhancement method with color restoration. [2] proposed a sand dust image reconstruction benchmark for training convolutional neural networks and evaluating the algorithm's performance, using the existing Pix2Pix network to restore sand dust images. However, the CNN-based architecture only considers the local features of the image, making the overall color projection problem of the image difficult to solve. At the same time, there are certain limitations in capturing long-range dependencies and recovering weak texture details. Based on this, vision transformer (ViT) [3, 4] came into being. ViT demonstrated the advantage of global processing and achieved a significant performance boost over CNN. Transformer can provide long-distance feature dependencies via the cascaded self-attention. However, it lacks the capability of retaining local feature details, thus leading to ambiguous and coarse details for image reconstruction. Therefore, it is very important to effectively combine CNN and Transfoemr.

Motivated by this, we propose a new design that brings together the power of Transformer and CNN into sand image restoration. The main idea is illustrated in Figure 1. Specifically, we first introduce an effective hybrid architecture that takes advantage of CNN and recent ViT for sand image restoration. We propose two branches (i.e., CNN and transformer branches) and aggregate them several times during the image restoration procedure. Consequently, local features extracted from the CNN branch and long-range dependencies captured in the transformer branch are progressively fused to complement each other and extract rich features. Experiments and comparisons demonstrate the superiority of our method over state-of-the-art sand image restoration methods.

In summary, our contributions are presented as follows:

- In comparison to pure CNN-based sand image restoration networks, our work is the first to introduce the power of Transformer into sand image restoration via novel designs.

- We propose a new fusion method to fuse CNN-based features with Transformer backbone-based features effectively.

- Extensive experiments on SandPascal VOC++ and real

Fig. 1. Overview structure of our method.

image datasets demonstrate the excellent performance of our method.

2. METHOD

In this section, we detail how to leverage the respective strengths of CNN and Transformer to promote their property in image restoration tasks to restore sharper images. Sand-Former consists of a shallow feature extraction module, a transformer branch, a CNN branch, and a high-quality projected image restoration module. The overall structure of SandFormer is shown in Figure 1.

2.1. Sand Images Formulation

The physical model that is widely used to describe the formation of an image suffered from light transmission hazed [5, 6] is often defined as follows:

$$I(x) = J(x)t(x) + A(1 - t(x)), \quad (1)$$

where $I(x)$ is the observed hazy image, A is the global atmosphere light, and $t(x)$ is the medium transmission map, $J(x)$ is the haze-free image. Moreover, we have $t(x) = e^{-\beta d(x)}$ being the atmosphere scattering parameter and the scene depth, respectively. Based on the physical model, we construct a new dataset called SandPascal VOC ++, which contains three forms of sand.

Dataset generation. In dusty weather, due to the different decay of R, G, and B values, the degraded images have prior features such as offset, concentration and time. Considering the atmospheric light attenuation effect of the dust floating in the atmosphere on the R, G, and B channels, the atmospheric light model belonging to the dust image was reconstructed according to the spatial distribution law. The mathematical expression of this model is:

$$\hat{A} = <A_R, k_1 A_R + b_1, k_2 A_R + b_2>, \quad (2)$$

Fig. 2. Analysis of the influence of different parameters in atmosphere scattering model.

where $A_G = k_1 A_R + b_1$, $A_B = k_2 A_R + b_2$, $\hat{A}$ is the global color deviation value of the sand dust image, k is the spatial distribution coefficient of the atmospheric light value of the three basic color spectrums, b is the disturbance amount. Based on the formula 2, various dust images with different degrees of degradation can be synthesized from clear images through artificial algorithms.

2.2. CNN Block

Although CNN has achieved great success in the field of image restoration, the inherent properties of CNNs make the network's performance reach a bottleneck. Based on this, this paper introduces ViT to enable the network to achieve better generalization performance while maintaining the ideal characteristics of CNNs. Motivated by this, we propose a novel dual-branch gated attention residual module to obtain local feature information. This module is implemented by embed-

Fig. 3. CNN Block

ding simple gating and channel attention in a convolutional neural network. As shown in Figure 3, the degraded image is obtained through ResNet[7] to obtain shallow network features of size $H_0 \times W_0 \times C_0$, which are respectively sent to the CNN branch and the Transformer branch. In order to keep the CNN features optimized in the CNN backbone, we add a downsampling module before each CNN Block to prevent overfitting. Both branches are deep-wise CNN to reduce the computational complexity; the convolution kernel sizes are 1 and 3, respectively, to achieve multi-scale feature extraction. To obtain a larger receptive field, we use simple channel attention at the end of each convolutional block, which fuses features at different scales to obtain richer feature representations.

2.3. Transformer Block

Fig. 4. Transformer Block

Due to the inductive bias of locality and weight sharing, the convolution operations demonstrate the intrinsic limitations in modeling the long-range dependency. The self-attention mechanism of Transfoemr itself makes it beneficial to capture long-term dependencies between input sequences. In order to take advantage of the powerful representation ability of Transformer, we add the Transformer branch to the proposed method and propose a novel feed forward network (FFN) called Dual-path Shared Weight Attention

FFN(DSWAFFN). Similar to the CNN branch, a dual-path form is also used to improve the fitting ability of this module. Meanwhile, to reduce the network complexity[8], the two branches share the weight, and the sigmoid activation function is added to obtain the attention weight of each branch. For the self-attention part of the transformer branch, we use the same structure as in[9], due to this can efficiently process high-resolution images while taking into account the capacity to handle global dependencies. A residual connection is at the end of this module, as shown in Figure 4, which allows the gradient to propagate effectively during backpropagation, thus avoiding the gradient-vanishing problem. Transformer stem aims to provide further guidance for global restoration with progressive features according to the convolution features.

2.4. Gate Fusion

Using CNN or Transformer separately causes either local or global features to be neglected, which affects the model's performance. We propose a novel fusion module to fuse features from different branches by gating blocks to address the feature inconsistency between Transformer and CNN. Specifically, feature maps are extracted from different levels of the CNN and Transformer branches and sent to the gate fusion module. We take a new step towards bridging the gap between CNNs and Transformer by presenting a new method to "softly" introduce a convolutional inductive bias into the ViT.

As shown in Figure 5, we also integrate the idea of gating into the design of fusion module. A simple gate block is added to replace the ReLU activation function in the middle part of the residual block. At the end of the fusion module, we use the channel dimension convolution block for further feature extraction. Using the fusion module, we can ingeniously integrate the features extracted by CNN with the features extracted by Transformer rather than directly adding them together, which significantly reduces the performance of the whole model.

3. EXPERIMENTS

3.1. Experimental results

This paper compares the sand dust image restoration method based on traditional prior and the haze image restoration method based on deep learning. For fairness, we retrain all networks on SandPascal VOC++. For the comparative experiment section, we have provided some visual effect comparison pictures in 6 and 7. We obtain the best results of the proposed method by training SandPascal VOC ++, combined with our proposed CNN branch, Transformer branch, and fusion module. We compare it with twelve State-of-the-Art methods and re-evaluate all methods, as shown in Table 1. Meanwhile, the visualized images in 6 and 7 match well with the quantitative results, showing our proposed method's favorable image restoration capability.

Fig. 5. We concatenate features from two branches, bidirectionally transfer the mixed information to the original branches.

Fig. 6. Image restoration results on a synthetic sand dust dataset.

Fig. 7. Image restoration results on the real world.

3.2. Ablation Study

For the ablation experiments section, as shown in Table 2, we illustrate the importance of individual components of our model. The model's performance is greatly reduced when there are only Transformer and CNN branches. Meanwhile, when the output of these two is simply added, the model's performance will also be affected. Finally, the three parts we proposed are integrated. That is, our proposed method achieves the best results, which shows that all parts of the network are of importance.

4. CONCLUSION

In this work, we explore the visual effects of SandFormer, formulating a synthetic dataset with simultaneously fugitive

Table 1. Comparative results on synthetic images and real images, all models are trained on our proposed dataset Sand-Pascal VOC++.

Method	SandPascal VOC++		Real		
	PSNR↑	SSIM↑	NIQE↓	NIMA↑	User Study↑
DCP[10]	17.925	0.855	3.361	3.998	2.6
MMSP[11]	17.772	0.843	2.926	3.943	7.3
GDCP[12]	13.669	0.752	3.255	3.890	3.9
TTFIO[13]	15.324	0.789	3.476	3.924	4.3
HRDCP[14]	12.212	0.683	4.131	3.884	5.8
OCM-GAT[15]	16.001	0.800	3.020	3.850	7.5
AOD-Net[16]	17.799	0.846	3.175	3.998	1.6
MSBDN[17]	29.033	0.918	2.988	3.726	3.6
4KDehazing[18]	28.215	0.912	3.447	3.852	5.3
AECR-Net[19]	27.222	0.907	3.146	3.886	4.2
Dehazeformer[20]	30.123	0.929	2.913	3.923	5.8
Transweather[21]	30.621	0.928	3.019	4.040	4.4
Sand_images	-	-	-	-	-
Ours	34.150	0.952	2.795	4.078	8.4

Table 2. Ablation experiments of proposed method.

	PSNR↑	SSIM↑
only Transformer Branch	31.426	0.941
only CNN Branch	20.449	0.826
Transformer + CNN	30.986	0.941
Trans. + CNN + SK Fusion	31.667	0.948
Trans. + CNN + Gate Fusion	34.150	0.952

dust, sand, and sandstorms. And proposed a sand dust image restoration network based on CNN and Transformer (called SandFormer), which combines the advantages of CNN in local feature recovery and Transformer in global perception, and obtains the final clear image through gate fusion. Extensive experiments show that our method outperforms state-of-the-art sand dust image restoration methods both quantitatively and qualitatively in dust scenes. In the future, we will incorporate high-level vision tasks.

5. REFERENCES

[1] Zhenghao Shi, Chunyue Liu, Wenqi Ren, Du Shuangli, and minghua Zhao, "Convolutional neural networks for sand dust image color restoration and visibility enhancement," *Chinese Journal of Image and Graphics*, vol. 27, no. 5, pp. 1493–1508, 2022.

[2] Yazhong Si, Fan Yang, Ya Guo, Wei Zhang, and Yipu Yang, "A comprehensive benchmark analysis for sand dust image reconstruction," *arXiv preprint arXiv:2202.03031*, 2022.

[3] 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.

[4] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10347–10357.

[5] Srinivasa G Narasimhan and Shree K Nayar, "Vision and the atmosphere," *International journal of computer vision*, vol. 48, no. 3, pp. 233–254, 2002.

[6] Yu Li, Shaodi You, Michael S Brown, and Robby T Tan, "Haze visibility enhancement: A survey and quantitative benchmarking," *Computer Vision and Image Understanding*, vol. 165, pp. 1–16, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[9] 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*, 2022, pp. 5728–5739.

[10] Kaiming He, Jian Sun, and Xiaoou Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011.

[11] Xueyang Fu, Yue Huang, Delu Zeng, Xiao-Ping Zhang, and Xinghao Ding, "A fusion-based enhancing approach for single sandstorm image," in *2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2014, pp. 1–5.

[12] Yan-Tsung Peng, Keming Cao, and Pamela C Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2856–2868, 2018.

[13] Zohair Al-Ameen, "Visibility enhancement for images captured in dusty weather via tuned tri-threshold fuzzy intensification operators," *International Journal of Intelligent Systems and Applications*, vol. 8, no. 8, pp. 10, 2016.

[14] Zhenghao Shi, Yaning Feng, Minghua Zhao, Erhu Zhang, and Lifeng He, "Let you see in sand dust weather: A method based on halo-reduced dark channel prior dehazing for sand-dust image enhancement," *Ieee Access*, vol. 7, pp. 116722–116733, 2019.

[15] Yan Yang, Chen Zhang, Longlong Liu, Gaoke Chen, and Hui Yue, "Visibility restoration of single image captured in dust and haze weather conditions," *Multidimensional Systems and Signal Processing*, vol. 31, no. 2, pp. 619–633, 2020.

[16] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng, "An all-in-one network for dehazing and beyond," *arXiv preprint arXiv:1707.06543*, 2017.

[17] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2157–2167.

[18] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Xiaobin Hu, Tao Wang, Fenglong Song, and Xiuyi Jia, "Ultra-high-definition image dehazing via multi-guided bilateral learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16180–16189.

[19] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma, "Contrastive learning for compact single image dehazing," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10546–10555.

[20] Yuda Song, Zhuqing He, Hui Qian, and Xin Du, "Vision transformers for single image dehazing," *arXiv preprint arXiv:2204.03883*, 2022.

[21] Jmj Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," *arXiv e-prints*, 2021.