A NOVEL WINDOW-INTERACTION MODULE BASED ON W-MSA

Ruochen Cui, Yongding Tao, Mingjun Ni *
University of Electronic Science and Technology of China
{ruochencui421, yongd.tao}@gmail.com, zomxsrin@163.com

ABSTRACT

W-MSA proposed by Swin Transformer has limitations in facilitating information interaction between windows. To address this, we introduce a module that utilizes convolution to achieve inter-window information interaction across different regions. Experiments demonstrate that our proposed module, when combined with W-MSA in a dual-branch structure, outperforms the simple W-MSA. In the deraining task conducted on the Uformer, we observe a 0.14dB improvement in performance. Our code: https://github.com/421zuoduan/WIM-code.

1 INTRODUCTION AND RELATED WORK

For a prolonged period, the field of computer vision, inclusive of tasks like derain, has been dominated by Convolutional Neural Networks (CNN)(Cheng et al. (2021); Zamir et al. (2021); Guo et al. (2021)). On the other hand, the evolution of network architectures in Natural Language Processing (NLP) takes a divergent route. Transformer(Vaswani et al. (2017)) emerges as a new paradigm exhibiting excellent performance in sequence modeling(Radford et al. (2018); Devlin et al. (2018); Rahali & Akhloufi (2023)). Its tremendous success in the realm of language prompts researchers to explore its adaptability to computer vision, giving rise to the Vision Transformer (ViT)(Dosovitskiy et al. (2020)). Compared to convolutional networks, ViT achieves an impressive balance between speed and accuracy in image classification.

In the domain of image restoration tasks, the significance of local information becomes exceptionally crucial. In contrast to ViT's global self-attention, Swin Transformer(Liu et al. (2021); Liang et al. (2021)) has accomplished self-attention within windows through window operations, prioritizing the extraction of local information, which is named as **W-MSA***(window based multi-head self-attention)*, thus achieving commendable outcomes. Moreover, to facilitate interaction between windowed information, Swin Transformer introduces shifting window operations.

However, their research paper's Ablation Experiment indicates that the enhancements from sliding windows were not markedly substantial. Subsequently, Uformer(Wang et al. (2022)), building upon the Swin Transformer, leverages the hierarchical structure of UNet(Ronneberger et al. (2015)) to establish multi-scale information and it achieves state-of-the-art results in derain tasks, enabling an alternative form of capturing local information. Nevertheless, Uformer still applied Swin Transformer's shifting window mechanism for inter-window information exchange. Due to this, we proposed a module based on window operations. This module generates convolutional kernels from local information and applies them globally to enable inter-window information exchange. Our experiments demonstrate improvements in derain tasks.

2 METHODOLOGY

Similar to the window partitioning mechanism of Swin Transformer, we initially divide the feature map into several windows. Here, we decompose a feature map of size $C \times H \times W$ into N^2 windows, each of size $C \times \frac{H}{N} \times \frac{W}{N}$, where N^2 denotes the number of windows, $\frac{H}{N}$ and $\frac{W}{N}$ represent the window size. Subsequently, we perform feature extraction by using the convolution operation of $k \times k$ for N^2 times.

*Corresponding Author

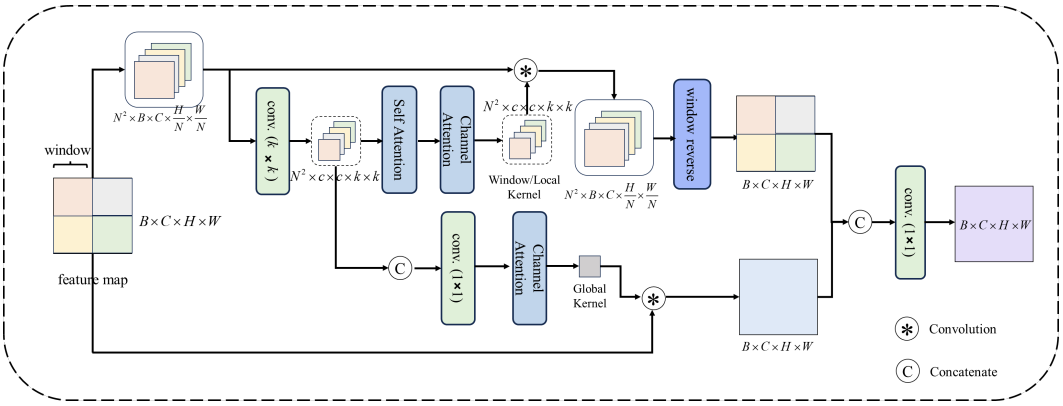

Figure 1: Window-Interaction Module

Furthermore, we get the convolutional kernels of size $C \times C \times k \times k$ and concatenate them along the output channel dimension. We then obtain a convolutional kernel with global information through a 1×1 convolution, and subsequently convolve it with the original feature map.

Next, we flatten the convolutional kernels obtained from the window convolution process, and concatenate them spacially. Calculate the spatial self-attention and then employ a SE module(Hu et al. (2018)) to compute channel-wise attention. After the attention calculation, the convolutional kernels are applied again to the original windows. Then, we use the regular window reverse to form a feature map. This operation bears similarity to the window reverse process employed by Swin Transformer.

Through the aforementioned operations, we obtain two feature maps of the same size as the input. We concatenate them along the channel dimension, and then reduce the channel dimension through a 1×1 convolution to obtain the output of the module. The detailed flow of the entire module is illustrated in Figure 1. The module above, in conjunction with W-MSA, collectively replaces the basic W-MSA. The outputs of both are concatenated along the channel dimension to obtain the final output.

3 Experiment

The dataset employed in this study is Rain100L(Yang et al. (2019)), which comprises 200 image pairs in the training set and 100 image pairs in the testing set. We opt for this dataset due to limitations in our available resources. Our code adopts the framework proposed in a referenced paper(Wu et al. (2022)).

Table 1: Results on the Rain100L dataset for single image derain

Method	PSNR	SSIM	Params	FLOPs
Uformer	38.680	0.97897	20.628M	10.308G
Ours	**38.820**	**0.97947**	24.667M	13.548G

Table 1 illustrates that our model outperforms our baseline Uformer by 0.14dB on the Rain100L dataset. The ablation study will be presented in the appendix, as illustrated in Table 2. The visualizations of the results can also be found in the appendix.

4 Conclusion

Overall, we conduct an analysis of the W-MSA proposed by Swin Transformer and identify areas for improvement. Building upon the W-MSA, we introduce a module that achieves inter-window information interaction through convolution. This module facilitates information exchange between windows through interactions among convolutional kernels. Experiments on Uformer demonstrate that this module contributes to performance improvement in deraining tasks.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4896–4906, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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.

Qing Guo, Jingyang Sun, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Wei Feng, Yang Liu, and Jianjun Zhao. Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1487–1495, 2021.

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

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*, pp. 1833–1844, 2021.

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*, pp. 10012–10022, 2021.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Abir Rahali and Moulay A Akhloufi. End-to-end transformer-based models in textual-based nlp. *AI*, 4(1):54–110, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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, 2015.

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.

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 (CVPR)*, pp. 17683–17693, June 2022.

Xiao Wu, T Huang, L Deng, and T Zhang. A decoder-free transformer-like architecture for high-efficiency single image deraining. In *Proc. IJCAI*, pp. 80, 2022.

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*, 42(6):1377–1393, 2019.

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*, pp. 14821–14831, 2021.

5 APPENDIX

5.1 ABLATION STUDY

We show our ablation study experiments in Table 2.From the table, we can know that our network works the best. Meanwhile, we can see that global kernel works as the most important part in our network from over 0.14dB difference between method nsw. and our network, with the average difference between methods containing global kernel and our network is 0.049dB.

Table 2: Ablation study of experiments of 1000 epochs on Rain100L. glok. : networks contains the global kernel. nse. : networks without the SE-Block of global kernel. nsw. : networks without shifted windows of Self-Attention Block. nsa. : networks without Self-Attention Block. nsep. : networks without SE-Block of previous kernels.

Method	PSNR	SSIM
glok.	38.718	0.97864
glok.+nse.	38.778	**0.97959**
glok.+nsw.	38.817	0.97939
nsw.	38.680	0.97879
nsa.+nsep.	38.787	0.97930
Ours	**38.820**	0.97947

5.2 STRUCTURE WITH FEWER PARAMETERS

For the issue of parameter increasing, we replace regular convolutions with depth-wise convolutions and remove the self-attention module while retaining the main structure, thus significantly reducing the parameters (-4.25M) and computational cost (-3.73FLOPs).

Table 3: Results on the Rain100L dataset for single image derain

Method	PSNR	SSIM	Params	FLOPs
Uformer	38.680	0.97897	20.628M	10.308G
Ours-L	**38.820**	**0.97947**	24.667M	13.548G
Ours-T	38.755	0.97909	**20.416M**	**9.82G**

In Table 3, "Ours-L" denote the model proposed in the main text, and "Ours-T" denote the model with less parameters which is proposed in the appendix. Table 3 illustrates that with less parameters, the model still brings performance improvement relative to the baseline, demonstrating the effectiveness of our proposed structure and expanding its applicability.

5.3 EXPERIMENT ON RAIN200L

Regarding the issue of a single and small dataset, we conducted additional experiments on a larger dataset, Rain200L (1800 pairs for training and 200 for testing, widely used in derain tasks).

Table 4: Results on the Rain200L dataset for single image derain

Method	PSNR	SSIM	Params	FLOPs
Uformer	40.751	0.98696	20.628M	10.308G
Ours-L	40.637	0.98671	24.667M	13.548G
Ours-T	40.708	0.98668	**20.416M**	**9.82G**

On Rain200L, as shown in Table 4, the model after structural adjustments achieved a performance nearly equivalent to the baseline, with a decrease of only 0.043dB in PSNR, while reducing the parameters (-0.21M) and computational cost (-0.49FLOPs), demonstrating comparable effectiveness to the baseline.

5.4 DISPLAY OF RAIN REMOVAL EFFECTS IN IMAGES

We show some images we tested in the dataset Rain100L and the details of them in Figure 2. From these processed images, we can see cleaner images and less rain streaks in our proceeding images compared with Uformer.

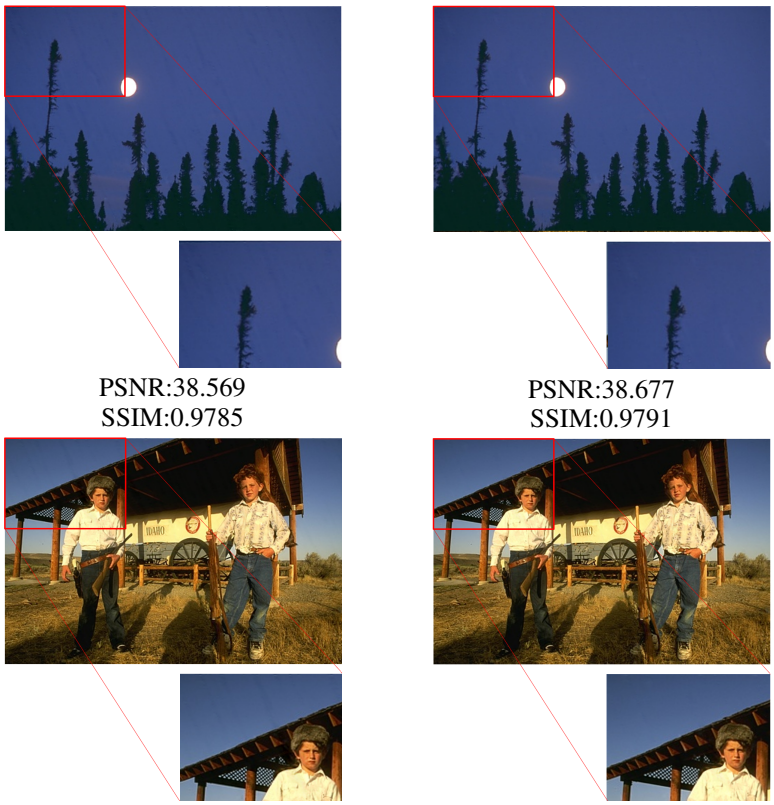

Figure 2: Some test images proceeded by Uformer and our network