A TRANSFORMER-BASED MOTION DEBLURRING NETWORK FOR UAV IMAGES

Rui Li, Xiaowei Zhao *

Intelligent Control & Smart Energy (ICSE) Research Group,
School of Engineering, University of Warwick, Coventry, CV4 7AL, UK
*Corresponding author: Xiaowei.Zhao@warwick.ac.uk

ABSTRACT

When performing surveying and mapping missions using drones, motion blur is an unavoidable issue induced by several factors such as vibration, turbulence and wind during operation. Such blurring can significantly degrade the image quality, adversely affecting the accuracy and reliability of downstream applications. In this paper, to effectively eliminate the motion blur in the UAV-captured images, we propose the NAFormer based on the well-established Nonlinear Activation Free Network (NAFNet), which introduces Transformer-based blocks to further enhance its motion-deblurring ability to UAV images. The experiments based on the UAVid dataset demonstrate the effectiveness of the proposed framework.

***Index Terms*—** UAV, Transformer, Motion Deblurring

1. INTRODUCTION

The continuous improvements in both hardware reliability and control strategy accelerate the practical applications of Unmanned Aerial Vehicles (UAV) in more and more areas [1] such as structural health monitoring [2] and precise crop mapping [3]. In actual operations especially for large-scale applications, the path-planning strategy usually tends to choose the shortest path, thereby minimizing operational time. Consequently, motion blurs will frequently and unavoidably occur in rapidly moving UAVs stem from environmental factors such as vibration, turbulence and wind. The motion blurs will definitely deteriorate the imaging quality, negatively impacting the downstream applications [4]. A direct yet effective solution is to actively detect and remove blurry images [5]. However, in certain specific applications such as 3D reconstruction, image removal may lead to an incomplete reconstruction result. Therefore, image deblurring, i.e. recovering a sharp image from a blurred input image, is a more versatile option.

As a classic problem in low-level computer vision, a series of deblurring methods have been proposed in the community. By assuming spatially uniform motion, the early deblurring methods model the blurs as the convolution of a blur kernel with a sharp image [6], thereby forming the kernel-based motion deblurring approaches employing regularized deconvolution. To effectively capture the intricate motion behaviours responsible for complex motion-induced blurs, it is often necessary to employ pixel or patch-wise motion flows. In pursuit of this goal, optical flow is frequently computed to facilitate the establishment of photometric constancy [7] between the restored latent sharp images, thereby paving the way for flow-based motion-deblurring methodologies. To address the temporal ordering ambiguity and the loss of spatial textures caused by motion blur, various priors have proven effective for blur kernels such as Gaussian scale mixture [8] and dark channel [9]. However, these priors may inadvertently lead to artefacts and compromise deblurring performance if the underlying assumptions are not met.

A different strategy is to directly train a neural network for motion deblurring, which frequently yields notable performance improvements. For example, Sun et al. [10] presented a CNN-based model to estimate a kernel and effectively eliminate non-uniform motion blur. Chakrabarti [11] employed a network to calculate estimations of sharp images affected by an unknown motion kernel. Additionally, Nah et al. [12] introduced a multi-scale loss function, implementing a coarse-to-fine strategy along with an adversarial loss for enhanced deblurring performance. In [13], Chen et al. proposed an efficient and effective Nonlinear Activation Free Network (NAFNet) following the UNet structure, which replaced or removed the nonlinear activation functions in the network. In this paper, we design a linear-attention-based [14, 15] Transformer block and incorporate it with the well-established NAFNet, achieving better performance with an insignificant complexity increase.

2. METHODOLOGY

The overall structure of the proposed NAFormer is shown in Fig. 1. Following the NAFNet [13], the NAFormer is built based on the UNet structure. Specifically, the blurred input is first fed into a convolutional layer before being processed by encoders and decoders. Both the encoder and decoder are constructed based on the Nonlinear Activation Free Block (NAFBlock) [13]. Transformer blocks refine the feature maps

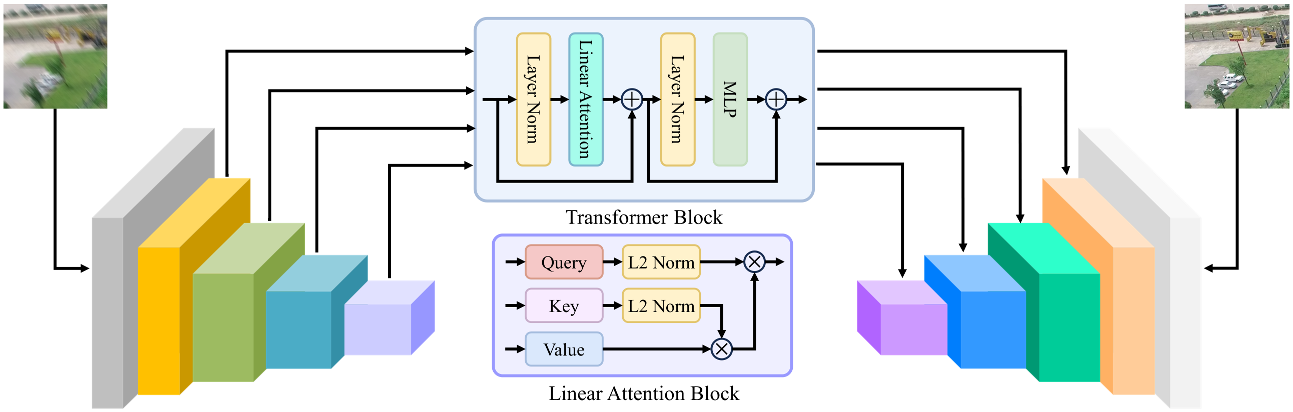

Fig. 1. The structure of the developed NAFormer.

generated by multi-scale encoders before being concatenated with feature maps obtained by corresponding decoders. Finally, the deblurred output is obtained after the final convolutional layer. As the details about the NAFBlock can be referred to [13], we mainly focus on introducing the developed Transformer block in this section.

Stemming from embedded self-attention mechanisms, the Transformer has strong capabilities to capture long-range dependencies and non-local relationships within the input. However, for self-attention mechanisms, an attention map needs to be generated to model the relationships between all pixel pairs in the input, whose shape is quadratically related to the input size. Therefore, when processing large-scale targets, both the time and space complexities of the Transformer will become too enormous to be used. In our previous work [14, 15], we decrease the complexity of self-attention mechanisms to linear by swapping the computational order. By taking advantage of the simplicity and practicality of linear attention, the coupling between Transformer blocks and multi-scale encoder-decoder structure becomes feasible. The details about the developed Transformer block and linear attention block can be seen in Fig. 1.

3. RESULTS

3.1. Dataset and metrics

To thoroughly evaluate the performance of the developed algorithm, we conduct experiments on the high-resolution UAVid dataset [16]. The UAVid dataset includes 420 images with 4K resolutions (3840 × 2160 or 4096 × 2160), mainly captured on urban street scenes by UAV.

To construct the training pairs, we randomly add the motion blur to raw images using the OpenCV library, while the kernel and angle of the blur are chosen from [5, 25] and $[-45°, 45°]$, respectively. To comprehensively evaluate the performance, two test sets are constructed, i.e. an easy set with an identical blurring setting with the training set and a hard set whose kernel and angle range from [11, 33] and $[-60°, 60°]$. The performance is compared by Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) in our experiments.

Table 1. Image deblurring results with different methods.

Test set	Model	PSNR	SSIM
Easy	Baseline	32.1473	0.9463
	NAFNet	32.7490	0.9530
	NAFormer	**32.8162**	**0.9535**
Hard	Baseline	27.1743	0.8487
	NAFNet	26.9030	0.8520
	NAFormer	**27.1923**	**0.8585**

Table 2. The comparison of complexity and inference speed between different methods. The results are generated based on the input in a shape of $3 \times 512 \times 512$.

Model	Complexity	Inference Speed
Baseline	63.46 GMacs	21.58 FPS
NAFNet	64.12 GMacs	20.42 FPS
NAFormer	75.97 GMacs	16.44 FPS

3.2. Experimental results

All models are trained for 200K iterations with the initial learning rate $1e^{-3}$ gradually reduced to $1e^{-6}$ with the cosine annealing schedule based on two RTX6000 GPUs. The training patch size is 512×512 and the batch size is 8. The complexity and inference speed are calculated based on input in size of $3 \times 512 \times 512$, on a single NVIDIA RTX6000 GPU.

As shown in Table 1, the proposed NAFormer achieve the best performance in both easy and hard test sets. Especially, the PSNR of the NAFNet experiences a slight decrease in the hard scenario, although the NAFNet outperforms the Baseline by a large margin in the easy test set. This may indicate a po-

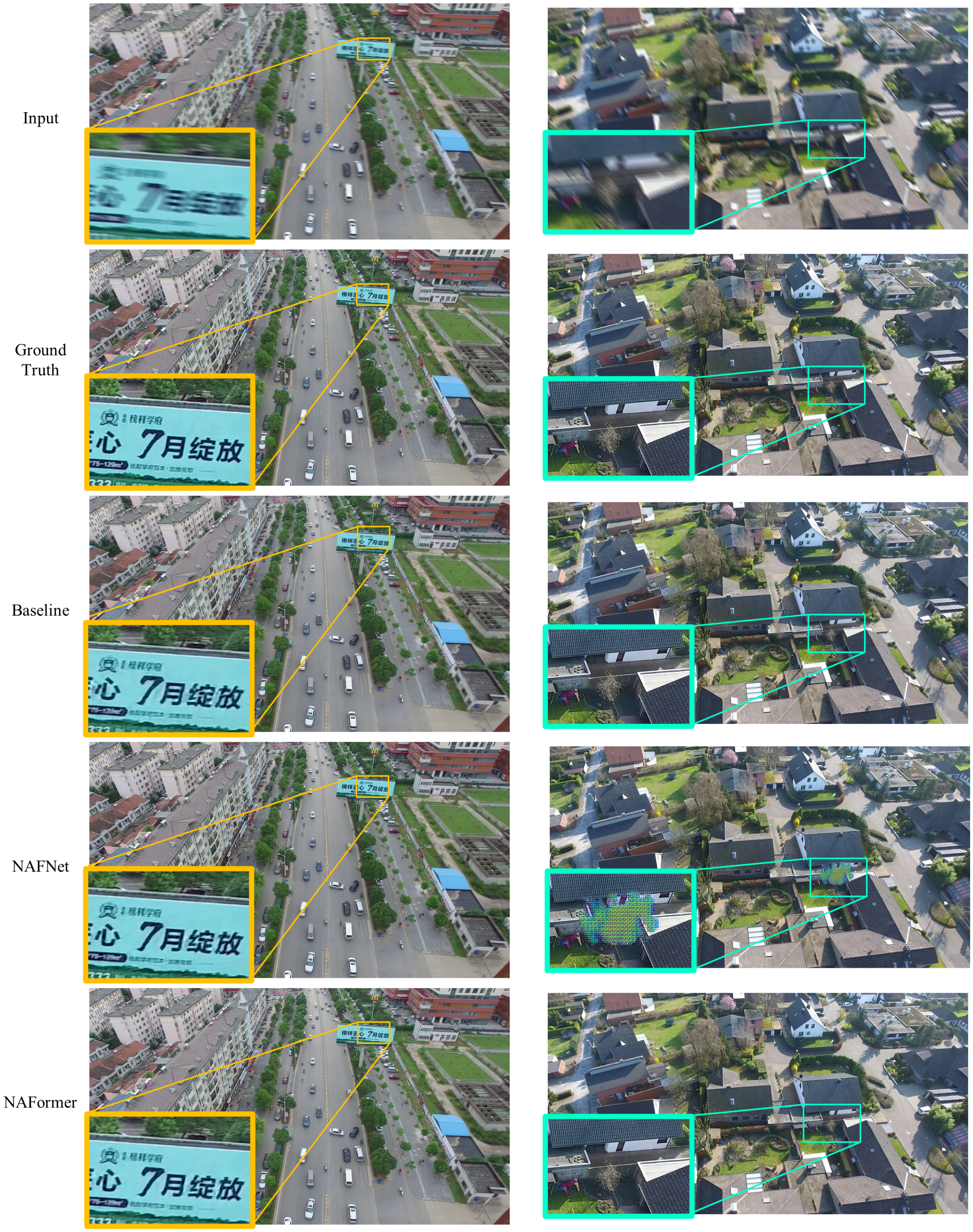

Fig. 2. Comparison of results generated by different methods for the (left) easy test set and (right) hard test set.

tential overfitting of the NAFNet for the training set. By contrast, the developed NAFormer maintains robust performance on both easy and hard test sets, demonstrating excellent generalizability. The visualization comparison in Fig. 2 shows a similar tendency. Taking the hard case as an example, certain texture details are totally lost by the NAFNet as shown in the enlarged part, which aligns with the above-mentioned performance decrease.

Besides performance, complexity and speed are also critical for evaluating a network, which is summarized in Table 2. Specifically, the NAFormer is about 20% slower than the NAFNet in terms of inference speed. Considering the stable performance and robust generalizability, we believe this level of speed reduction is acceptable.

4. CONCLUSION

In this paper, we developed the NAFormer for motion deblurring of UAV images, which integrated NAFNet and Transformer. The experiment results show that our NAFormer demonstrated a robust performance under different experimental settings. In our forthcoming research, we intend to assess the effectiveness of models pre-trained on synthetic datasets when applied to real-world UAV-captured images with motion blur. The developed workflow will be used for UAV-based offshore renewable energy infrastructure monitoring.

5. ACKNOWLEDGMENTS

The authors acknowledge the Scientific Computing Research Technology Platform (SCRTP) at the University of Warwick for providing HPC resources.

6. REFERENCES

[1] Q. Zhang, S. Zheng, C. Zhang, X. Wang, and R. Li, “Efficient large-scale oblique image matching based on cascade hashing and match data scheduling,” *Pattern Recognition*, vol. 138, p. 109442, 2023.

[2] S. Zhao, F. Kang, J. Li, and C. Ma, “Structural health monitoring and inspection of dams based on uav photogrammetry with image 3d reconstruction,” *Automation in Construction*, vol. 130, p. 103832, 2021.

[3] A. López, J. M. Jurado, C. J. Ogayar, and F. R. Feito, “A framework for registering uav-based imagery for crop-tracking in precision agriculture,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 97, p. 102274, 2021.

[4] F. Xu, L. Yu, B. Wang, W. Yang, G.-S. Xia, X. Jia, Z. Qiao, and J. Liu, “Motion deblurring with real events,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2583–2592.

[5] T. Sieberth, R. Wackrow, and J. H. Chandler, “Automatic detection of blurred images in uav image sets,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 122, pp. 1–16, 2016.

[6] D. Krishnan, T. Tay, and R. Fergus, “Blind deconvolution using a normalized sparsity measure,” in *CVPR 2011*. IEEE, 2011, pp. 233–240.

[7] T. Hyun Kim and K. Mu Lee, “Generalized video deblurring for dynamic scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5426–5434.

[8] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, “Removing camera shake from a single photograph,” in *Acm Siggraph 2006 Papers*, 2006, pp. 787–794.

[9] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, “Blind image deblurring using dark channel prior,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1628–1636.

[10] J. Sun, W. Cao, Z. Xu, and J. Ponce, “Learning a convolutional neural network for non-uniform motion blur removal,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 769–777.

[11] A. Chakrabarti, “A neural approach to blind motion deblurring,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 221–235.

[12] S. Nah, T. Hyun Kim, and K. Mu Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3883–3891.

[13] L. Chen, X. Chu, X. Zhang, and J. Sun, “Simple baselines for image restoration,” in *European Conference on Computer Vision*. Springer, 2022, pp. 17–33.

[14] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, “Abcnet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery,” *ISPRS journal of photogrammetry and remote sensing*, vol. 181, pp. 84–98, 2021.

[15] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, “Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[16] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, “Uavid: A semantic segmentation dataset for uav imagery,” *ISPRS journal of photogrammetry and remote sensing*, vol. 165, pp. 108–119, 2020.