BlazeBVD: Make Scale-Time Equalization Great Again for Blind Video Deflickering — ECCV 2024 Supplementary Materials

Xinmin Qiu¹, Congying Han¹, Zicheng Zhang¹, Bonan Li¹, Tiande Guo¹, Pingyu Wang², and Xuecheng Nie³

¹ University of Chinese Academy of Sciences, Beijing, China
 ² Sichuan University, Sichuan, China
 ³ MT Lab, Meitu Inc., Beijing, China

1 Overview

In this supplemental material, additional experimental details and results are provided, including:

- Details about the effect of TCM on video continuity (in Sec. 2);
- Details about network architecture (in Sec. 3);
- More qualitative comparisons with the baseline (in Sec. 4);
- Details about user studies (in Sec. 5);
- Future work (in Sec. 6);
- Video comparisons. Please refer to SupplementaryVideo.mp4.

2 Mutual Information

Here, we theoretically prove the effect of TCM on video continuity by calculating inter-frame mutual information (MI) [3]. Given the current frame X_t and the previous frame X_{t-1} , with the current frame adjusted by TUnet as \hat{X}_t , we can calculate the mutual information between X_t and X_{t-1} , and the mutual information between \hat{X}_t and X_{t-1} :

$$MI(X_{t-1}, X_t) = H(X_t) - H(X_t | X_{t-1})$$

$$= H(X_t) + \sum_{x_i^{t-1}} p(x_i^{t-1} \sum_{x_j^t} p(x_j^t | x_i^{t-1}) \log(p(x_j^t | x_i^{t-1})))$$

$$= H(X_t) + \frac{1}{HW} \sum_{x_i^{t-1}, x_j^t} p(x_j^t | x_i^{t-1}) \log p(x_j^t | x_i^{t-1}),$$

$$MI(X_{t-1}, \hat{X}_t) = H(\hat{X}_t) - H(\hat{X}_t | X_{t-1})$$

$$= H(\hat{X}_t) + \sum_{x_i^{t-1}} p(x_i^{t-1} \sum_{\hat{x}_j^t} p(\hat{x}_j^t | x_i^{t-1}) \log(p(\hat{x}_j^t | x_i^{t-1})))$$

$$= H(\hat{X}_t) + \frac{1}{HW} \sum_{x_i^{t-1}, \hat{x}_j^t} p(\hat{x}_j^t | x_i^{t-1}) \log p(\hat{x}_j^t | x_i^{t-1}),$$

$$(1)$$

where $H(X_t) = H(\hat{X}_t)$ and $p(\hat{x}_j^t | x_i^{t-1}) \ge p(x_j^t | x_i^{t-1})$. So we can get $MI(X_{t-1}, \hat{X}_t) \ge MI(X_{t-1}, X_t)$, which indicates that the inter-frame communication degree is higher.

3 Network Architecture

Table 1 shows the specific details of the network architecture in BlazeBVD, including the models of GFRM, LFRM and TCM. NonLocal denotes Non-local neural networks [2], which are used to search for inter-frame motion information.

| Module | Layer | Kernel Size | Channel | Layer | Kernel Size | Channel |
|--------|--------------------------------|--------------------|--|--------------------------------|--------------------|--|
| GFRM | encoder1 | $3 \times 3/(1,1)$ | $6 \rightarrow 32$ | pool1 | $2{	imes}2/(2,2)$ | $32 \rightarrow 32$ |
| | encoder2 | $3{	imes}3/(1,1)$ | $32 \rightarrow 64$ | pool2 | $2{	imes}2/(2,2)$ | $64 \rightarrow 64$ |
| | encoder3 | $3{	imes}3/(1,1)$ | $64 \rightarrow 128$ | pool3 | $2{	imes}2/(2,2)$ | $128 \rightarrow 128$ |
| | encoder4 | $3{	imes}3/(1,1)$ | $128 \rightarrow 256$ | pool4 | $2{	imes}2/(2,2)$ | $256 \rightarrow 256$ |
| | bottleneck | $3{	imes}3/(1,1)$ | $256 {\rightarrow} 512$ | - | - | - |
| | decoder4 | $3 \times 3/(1,1)$ | $512 \rightarrow 256$ | upconv4 | $3 \times 3/(1,1)$ | $512 \rightarrow 256$ |
| | decoder3 | $3 \times 3/(1,1)$ | $256 \rightarrow 128$ | upconv4 | $3 \times 3/(1,1)$ | $256 \rightarrow 128$ |
| | decoder2 | $3 \times 3/(1,1)$ | $128 \rightarrow 64$ | upconv4 | $3 \times 3/(1,1)$ | $128 \rightarrow 64$ |
| | decoder1 | $3 \times 3/(1,1)$ | $64 \rightarrow 32$ | upconv4 | $3 \times 3/(1,1)$ | $64 \rightarrow 32$ |
| | conv | $1 \times 1/(1,1)$ | $32 \rightarrow 3$ | - | - | - |
| LFRM | $\operatorname{conv} \times 2$ | $3 \times 3/(1,1)$ | $9 \rightarrow 32$ | NonLocal | - | $32 \rightarrow 32$ |
| | conv | $1 \times 1/(1,1)$ | $32 \rightarrow 32$ | NonLocal | - | $32 \rightarrow 3$ |
| TCM | $\operatorname{conv} \times 2$ | $3 \times 3/(1,1)$ | $16 \rightarrow 32 \rightarrow 64$ | - | - | - |
| | $Transformer \times 8$ | $3 \times 3/(1,1)$ | $64 \rightarrow 128 \rightarrow 64$ | bilinear | - | $64 \rightarrow 64$ |
| | $\operatorname{conv} \times 2$ | $3 \times 3/(1,1)$ | $64 \rightarrow 32 \rightarrow 16$ | conv | $3 \times 3/(1,1)$ | $3 \rightarrow 16$ |
| | $\operatorname{conv} \times 3$ | $3{	imes}3/(1,1)$ | $32 \rightarrow 8 \rightarrow 4 \rightarrow 1$ | $\operatorname{conv} \times 3$ | $3{	imes}3/(1,1)$ | $32 \rightarrow 16 \rightarrow 16 \rightarrow 3$ |

 Table 1: Detailed architecture of our BlazeBVD.

4 Qualitative Comparisons

We show more qualitative comparisons between BlazeBVD and the baseline on synthetic videos and DAVIS-2017-Test. As shown in Fig. 1 and Fig. 2, BlazeBVD maintains the color of moving objects more accurately compared to the color artifacts brought by Deflicker [1]. This is due to the fact that the atlas-based representation in Deflicker does not work well for moving objects.

5 Details about User Study

To better compare the video quality of BlazeBVD and baseline after flicker removal, we conduct a user study on Amazon Mechanical Turk following an



Fig. 1: Qualitative comparisons between the baseline and our BlazeBVD.



Fig. 2: Qualitative comparisons between the baseline and our BlazeBVD.

4 X. Qiu et al.

A/B test protocol in real-world videos and generation videos. Specifically, we randomly select 50 videos from seven datasets. Each user needs to choose a video with better perceptual quality and better fidelity about content from videos processed by our method and the baseline. In total, we have 50 users and 50 pairs of comparisons. Our method outperforms the baseline significantly, as shown in Fig. 3.



Fig. 3: Results of user study. A total of 50 users make a choice among 50 videos, of which randomly Expert contains 3 videos, OldAnime contains 8 videos, OldMovie contains 12 videos, SlowMotion contains 6 videos, TimeLapse contains 7 videos, VideoDM contains 8 videos, and VideoLDM contains 6 videos.

6 Future work

The proposed method can be applied to various flickering videos, which does not need to provide extra guidance, and can be directly applied to processed videos of other tasks for further refinement. During the deflickering process, we realized that the temporal consistency of the video content needed to be additionally considered. Especially in the videos obtained by the generative model, maintaining the fidelity of the content and improving the fluency are contradictory. How to better balance the degradation between faithfulness and coherence when processing videos and find reasonable metric forms is our future work.

References

- 1. Lei, C., Ren, X., Zhang, Z., Chen, Q.: Blind video deflickering by neural filtering with a flawed atlas. In: CVPR. pp. 10439–10448 (2023)
- Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
- Zhao, Y., Li, Z., Guo, X., Lu, Y.: Alignment-guided temporal attention for video action recognition. In: NeurIPS. vol. 35, pp. 13627–13639 (2022)