Event-guided Deblurring of Unknown Exposure Time Videos –Supplementary Material–

Taewoo Kim[1], Jeong-Min Lee[1], Lin Wang[2], and Kuk-jin Yoon[1]

[1] Korea Advanced Institute of Science and Technology
{intelpro, jeanmichel, kjyoon}@kaist.ac.kr
[2] AI Thrust, HKUST Guangzhou and Dept. of CSE, HKUST
linwang@ust.hk

Abstract. Due to the lack of space in the main paper, we provide more details of the proposed method and experimental results in the supplementary material. Sec.1 adds the related works for image reconstruction using the event camera and cross-modal attention. Sec.2 provides the detailed network architectures of the proposed RNN-cell for the event feature encoding. Lastly, Sec.3 presents the details of data collection and the experiments on dataset generations. Finally, Sec.4 presents additional experimental results and details.

1 Additional related work

Deep learning for Event-to-Image and Video Reconstruction. The other line of research directly reconstructs sharp images and video from event data via adversarial learning [22, 12, 29], RNNs [18, 31, 20], and self-supervised learning [16]. As the event cameras, *e.g.*, DAVIS 240C [1], are in a low-resolution, some attempts have tried to reconstruct high-resolution images via supervised [11] and unsupervised learning [25]. Moreover, some works [24, 23] have demonstrated that image reconstruction can be used to help event-based visual perception tasks in training. However, reconstructing video from the events is still a highly ill-posed problem due to inherently unstable contrast threshold and sensor noise.

Cross-Modal Attention Attention mechanisms can adaptively transform a network's parameters according to inputs. Thus, it boosts representative features while suppressing uninformative features in various manners, such as channel attention, spatial attention, and temporal attention [6, 21, 3, 17, 26, 9]. Recently, a growing body of research has been delving into dynamic feature modulation considering two modality inputs. For event and frame modalities, Gehrig *et al.* [5] introduce a recurrent feature modulation mechanism to fuse the event and RGB sensor data.

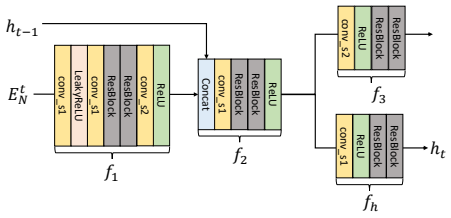


Fig. 1: The proposed RNN-cell for recurrent encoding of the events

Table 1: Dataset comparison between Blur-DVS and our datasets

	Camera	Resolution	Color	# sharp images	# Scenes
Blur-DVS [7, 10, 19]	DAVIS-240C	240×180	No	15246	-
Our datasets	Color DAVIS-346	346×260	Yes	53601	59

2 Network Architecture

2.1 Recurrent encoding for the embedded events

To apply the proposed RNN-cell, we first divided B temporal bins of the voxel grid of the events into N temporal units as mentioned in the main paper. For each temporal-wise divided unit event $\{E_t^n\} \in \mathbb{R}^{2\times H\times W}$ with temporal index $n \in \{1, ..., N\}$, we apply the proposed RNN cell as illustrated Fig.1. We first extract the feature of the unit event utilizing the first encoding block f_1 as follows:

$$\mathcal{F}(E_t^n)_{s=0} = f_1(E_t^n) \tag{1}$$

We then generate a feature map of the next scale.

$$\mathcal{F}(E_t^n)_{s=1} = f_2(Concat(\mathcal{F}(E_t^n)_{s=0}, h_{t-1})) \tag{2}$$

where $Concat$ denotes channel-wise concatenation operation; f_2 refers to the second CNN block of the RNN cell, and h_{t-1} refers to the previously generated hidden state. With these local event feature $\mathcal{F}(E_t^n)_{s=1}$, we recursively update hidden state h_t as follows:

$$h_t = f_h(\mathcal{F}(E_t^n)_{s=1}) \tag{3}$$

where f_h denotes the CNNs block for extracting the hidden state. We then further process $\mathcal{F}(E_t^n)_{s=1}$ using the last CNNs block f_3 represented as:

$$\mathcal{F}(E_t^n)_{s=2} = f_3(\mathcal{F}(E_t^n)_{s=1}) \tag{4}$$

In this way, we generate the output hierarchical feature maps $\{\mathcal{F}(E_t^1)_s, ..., \mathcal{F}(E_t^N)_s\}$ for the current part($s \in \{0, 1, 2\}$). All the generated feature maps are concatenated with the feature map of events for the past part.

Fig. 2: DAVIS-346 Color camera used for dataset collection.

3 Real-world event datasets

3.1 Dataset collection details and comparisons

As mentioned in the main paper, there are no publicly available large-scale datasets for evaluating event-guided motion deblurring, including real-world events. The previous event-guided motion deblurring methods used the dataset called named as Blur-DVS [7, 10, 19], which is not publicly available. For this reason, we collected a new dataset using the Color-DAVIS 346 camera that provides RGB images and spatially aligned stream of events as shown in Fig.2. Since the Color-DAVIS 346 camera has a low frame rate (maximum $\sim$ 40fps), we captured static scenes when collecting sharp images. We minimize motion blur by moving the camera slowly. Compared to Blur-DVS [7, 10, 19], we collected more sharp images and diverse scenes, as shown in Table.1. In addition, we obtained relatively higher resolution events and images pair since the Color-DAVIS 346 camera can shoot events and frames with a higher resolution than the DAVIS-240C. Finally, our dataset contains RGB images, whereas Blur-DVS [7, 10, 19] only provides intensity images. Therefore, we can evaluate richer texture information of the scene details using RGB images.

Table 2: The evaluation results of our model on the GoPro dataset when tested in different configurations from the training set using three different training datasets. We trained our network for the number of iterations the same as used in the main paper.

Training datasets	Unseen interval							
	7-5		9-3		11-1		Avg.	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Known exposure time	23.26	0.7354	25.03	0.7881	31.26	0.9263	26.52	0.8166
Unknown exposure time	37.39	0.9700	34.62	0.9541	34.94	0.9584	35.65	0.9608
Unknown exposure time + random noise	37.07	0.9686	36.84	0.9669	36.41	0.9642	36.77	0.9666

3.2 The experiments on dataset generations

For performing event-guided motion deblurring for unknown exposure time videos, it is crucial to simulate various unknown video frame acquisition processes during the training phase(arbitrary exposure and readout time). The reason is that there can be arbitrary different ratios for exposure and readout interval in real situations. Therefore, our ETES module needs to learn select event features corresponding to unknown exposure time at various arbitrary interval ratios. To this end, we add random noise to readout-interval for the generalization ability. To demonstrate the effectiveness of our proposed dataset generation method, we trained our model with three different training datasets and tested our model not seen in the training set. First, we train our model with the assumption of the previous event-guided motion deblurring methods(shutter period and exposure time are the same - known exposure assumption). In that case, the performance is dramatically degraded to unseen combination of exposure and read-out time, as shown in the first row of Table.2. Next, we generated a training sets with $m+n = 16$, set the number of video frames of the exposure phase m=$\{9,11,13,15\}$ as in the main paper. We then train our network without adding random noise to evaluate the performance. We observe the deblurring performance is somewhat improved in the unseen interval, but the still degraded performance(the 2nd row of Table.2). Lastly, we observe a significant performance improvement to the unseen interval by adding random noise as shown in the 3rd row of Table.2. This experiment demonstrated that we could improve the generalization ability in the unseen exposure-readout intervals by adding random noise to the readout interval.

4 Additional experimental results and details

4.1 Video results

Video results on GoPro-15fps datasets with the unknown and random exposure time of the frame-based camera To simulate a situation where the inconsistent exposure time(temporally varying), we arbitrarily change the number of frames corresponding to the exposure time(m in the main paper) in the GoPro-15fps dataset. In the same manner as the main paper, we set $m + n = 16$. For simulating random exposure time, we randomly select m among $\{9, 11, 13, 15\}$. Please note that we performed motion deblurring without exposure time information on random exposure time videos. Demo videos named video_results_gopro.mp4 for random exposure time videos includes input, ground truth, the predicted results of state-of-the-art(SoTA) video deblurring methods CDVD-TSP [14], and proposed our methods.

Video results on our real-world event datasets with the unknown and random exposure time of frame-based camera In the same manner as the GoPro-15fps dataset, we simulate the random exposure time video by randomly

selecting the frame number of exposure phase m. We set $m + n = 14$ and randomly choose m among $\{9, 11, 13\}$. Demo videos named video_results_dvs.mp4 includes input, ground truth, the predicted results of state-of-the-art(SoTA) event-guided video deblurring methods D2Nets [19], and proposed our method.

Video results on unknown exposure time real-world blurry videos Finally, we generated video demos of unknown exposure time real-world blurry videos. Demo videos named video_results_real_blur.mp4 includes input, the predicted results of state-of-the-art(SoTA) video deblurring methods CDVD-TSP [15], SoTA event-guided video deblurring methods D2Nets [19], and proposed our method.

4.2 Additional visual results

More visual comparisons on real-world unknown exposure time blurry video frames In Fig. 4 and Fig. 5 and Fig. 6, we perform qualitative comparisons with the SoTA frame-based image deblurring methods (MIMOUNet+ [4]) and the SoTA event-guided video deblurring methods(D2Nets [19]) on real-world blurry video frames captured by Color DAVIS-346 event camera. We confirm that our method restores more precise, sharp details than other methods, even in real-world blurry video frames.

More visual comparisons on the test split of our real-world event datasets In Fig. 7 and Fig. 8 and Fig. 9 and Fig. 10, we perform qualitative comparisons with the SoTA frame-based image deblurring methods(DMPHN [28], MIMOUNet+ [4], MPRNet [27]) and the SoTA video deblurring method(CDVD-TSP [14]) and the SoTA event-guided video deblurring method (D2Nets [19]). We confirm that our method can more precisely restore sharp images even in severe blurry conditions caused by non-linear motion.

More visual comparisons on GoPro-15fps dataset [13] In Fig. 11 and Fig. 12 and Fig. 13 and Fig. 14, we perform qualitative comparisons with the SoTA frame-based image deblurring methods(HINet [2], MIMOUNet+ [4], MPRNet [27]) and video deblurring methods(ESTRNN [30], CDVD-TSP [14]). Our proposed networks can restore a more plausible and sharp image than other methods.

4.3 Additional average temporal activation maps of our ETES modules on real-world blurry videos

In the main paper, we experimented with the ETES module's estimation results of unknown exposure time on real-world blurry videos. In addition, we experimented on the exposure time estimation result of our ETES module for motion

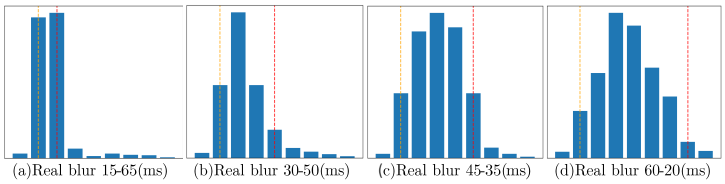


Fig. 3: The visualization results of the average temporal activation map of the ETES module on the real-world blurry videos. The horizontal and vertical axes represent the temporal axis and the average amount of channel activation, respectively. The yellow and red dotted lines indicate the start and end of the exposure time, respectively. The first and the last numbers indicate the exposure and readout time of real-world blurry videos.

deblurring with a different combination of the exposure time-readout time setting included in the main paper. For this experiment, we set exposure time as {15, 30, 45, 60}ms and shutter period as 80ms. We then plotted averaged temporal activation map of the ETES module for 200 video frames of each video clip in Fig.3. Here, we confirmed that all activation is hardly activated in the readout phase and mainly in the exposure phase, even in the different compositions of main paper.

4.4 Implementation details of other event-guided methods

We retrain other event-guided methods(LEDVDI† [10], DMPHN† [28], Nah *et al.*† [13], D2Nets† [19]) for same iterations as with our method. For all datasets, we utilize the batch size of 8 and ADAM [8] optimizer to update weight using a multi-step scheduler with an initial learning rate of 1×10^{-4} and decay rate of 0.5. For data augmentation, we apply random cropping(256×256) to the event and frame for the same position. Since D2Nets† [19] uses network input for ground truth sharp frame(non-consecutively blurry frames assumption), we replace ground truth sharp frame with blurry frame for fair comparisons. As can be seen in the case of LEDVDI† [10](Tab.3), the retrained model shows much better results than the official pretrained model.

Table 3: The evaluation results of LEDVDI† [10] on the real-world event dataset using the pretrained model and retrained model. Please note the performance of retrained model is much better than when using pretrained model.

	9-5		11-3		13-1		Avg.	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LEDVDI† [10](pretrained)	23.54	0.6783	23.44	0.6751	24.21	0.6986	23.73	0.6840
LEDVDI† [10](retrained)	34.77	0.9258	33.83	0.9138	32.96	0.9047	33.86	0.9148

Table 4: The ablations of network according to various scales.

	Real-world event dataset				Complexity FLOPs(G)
	9-5	11-3	13-1	Avg.	
	PSNR	PSNR	PSNR	PSNR	
Ours(2-scale)	36.36	35.46	35.74	35.85	149.79
Ours(3-scale)	**36.98**	**36.06**	**35.98**	**36.35**	237.77
Ours(4-scale)	<u>36.64</u>	<u>35.95</u>	<u>35.68</u>	<u>36.09</u>	324.89

4.5 Implementation details of other frame-based methods

As with the event-guided methods, we retrain frame-based image deblurring method(MPRNet [27], MiMOUNet+ [4], DMPHN [28], Nah *et al.* [13]) and video deblurring method (CDVD-TSP [14]) for 3.75×10^5 iterations in the real-world event dataset under their original hyperparameter setting provided by authors. We apply random cropping(256×256) to the frame during training. For all frame-based deblurring methods training, we used the official GitHub code provided by authors.

4.6 Scale-ablation study

In Tab.4, we report the performance of the network structure according to various scales. Using the 2-scale network structure reduces the cost; meanwhile, we observed a slight performance drop. We also observed a marginal performance variation with the 4-scale network. So, we selected the 3-scale network as our backbone structure when considering the trade-off between complexity and performance.

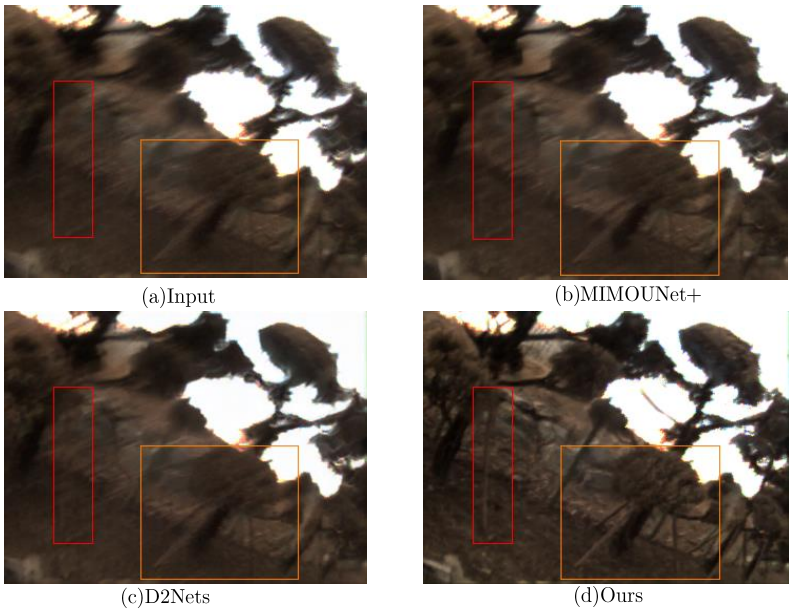


Fig. 4: Visual comparisons on **real-world** unknown exposure time blurry video frames.

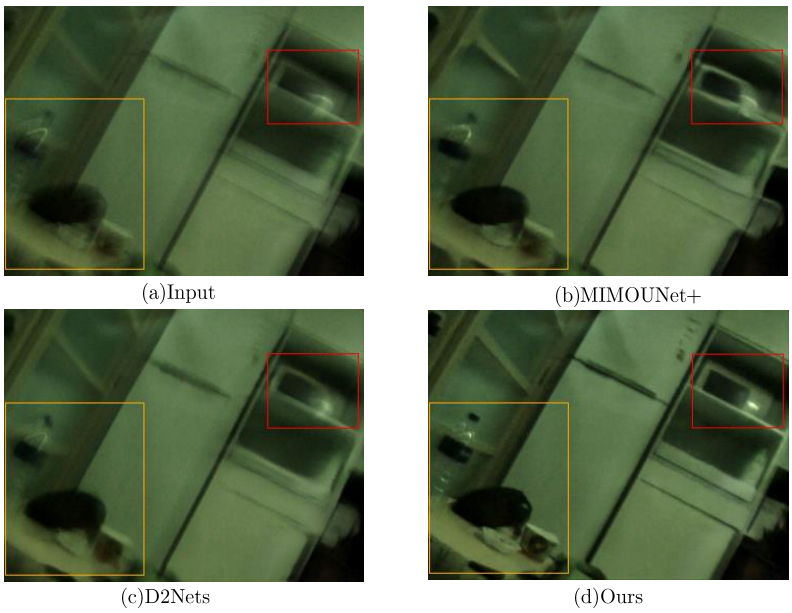
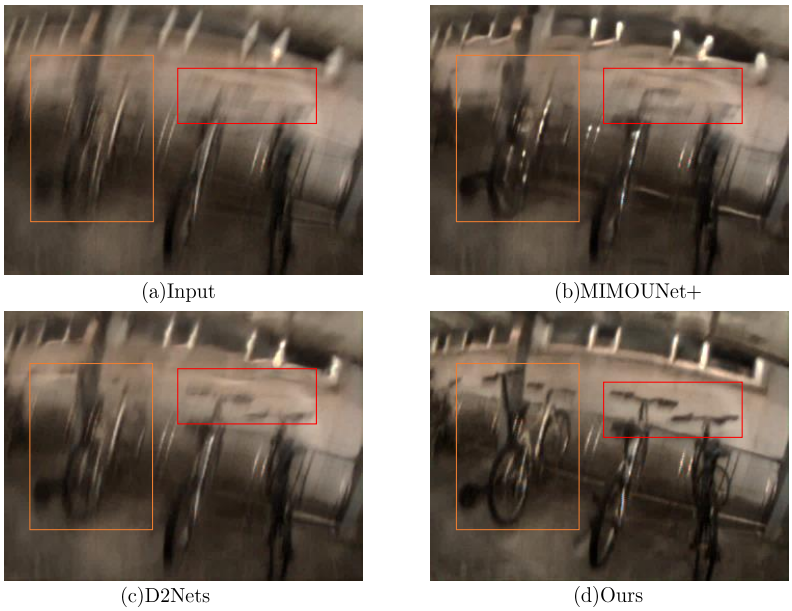


Fig. 5: Deblurring results on **real-world** unknown exposure time blurry video frames.



(a)Input (b)MIMOUNet+

(c)D2Nets (d)Ours

Fig. 6: Deblurring results on **real-world** unknown exposure time blurry video frames.

Fig. 7: Visual comparison of unknown exposure time blurry video frames on our real-world event datasets.

Fig. 8: Visual comparison of unknown exposure time blurry video frames on our real-world event datasets.

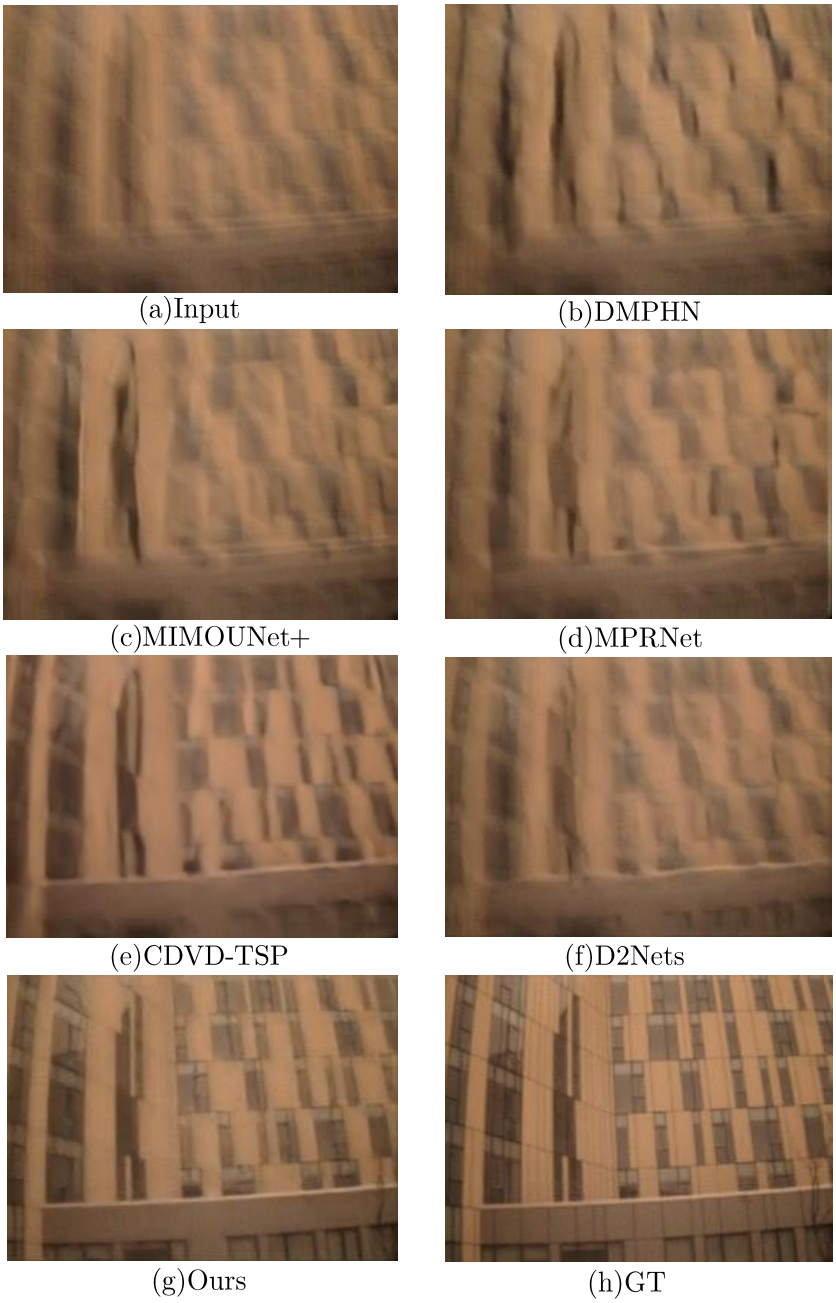


Fig. 9: Visual comparison of unknown exposure time blurry video frames on our real-world event datasets.

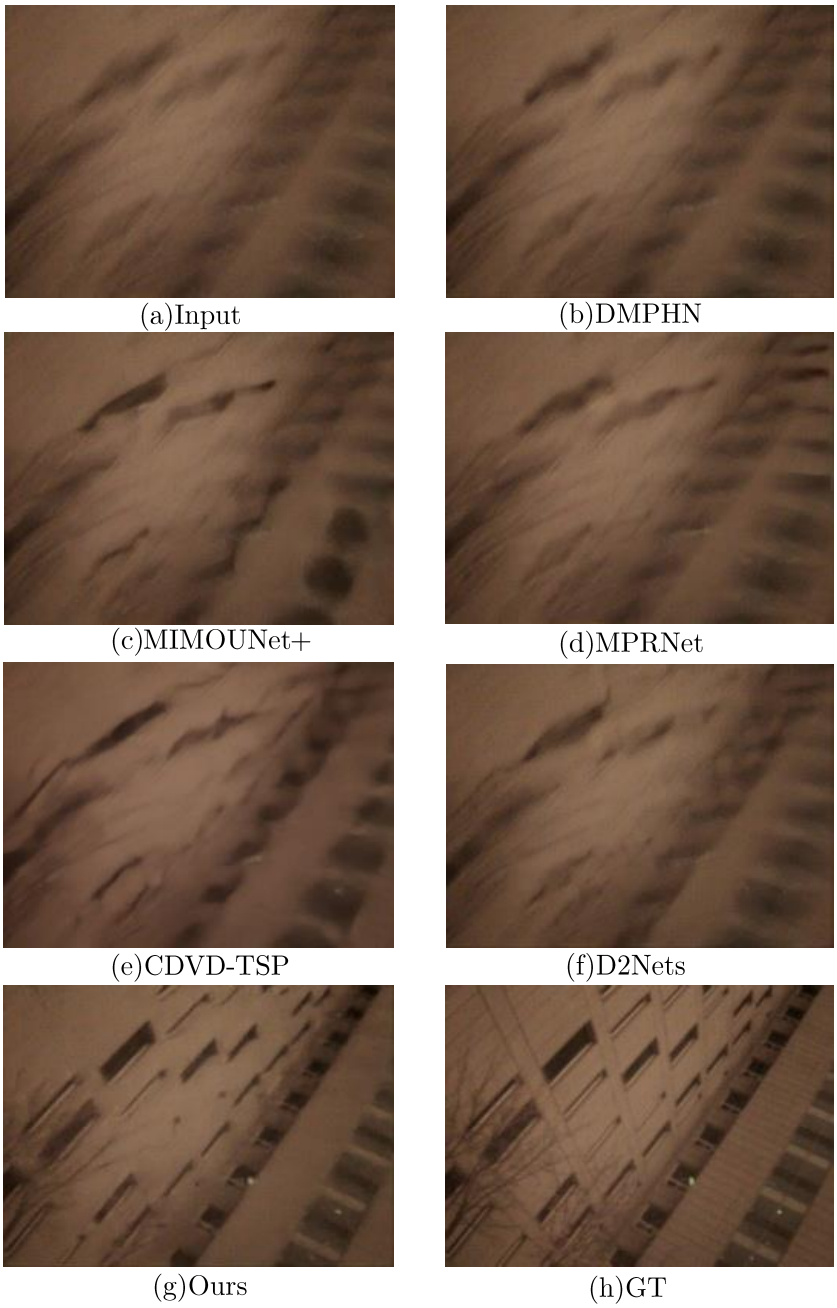
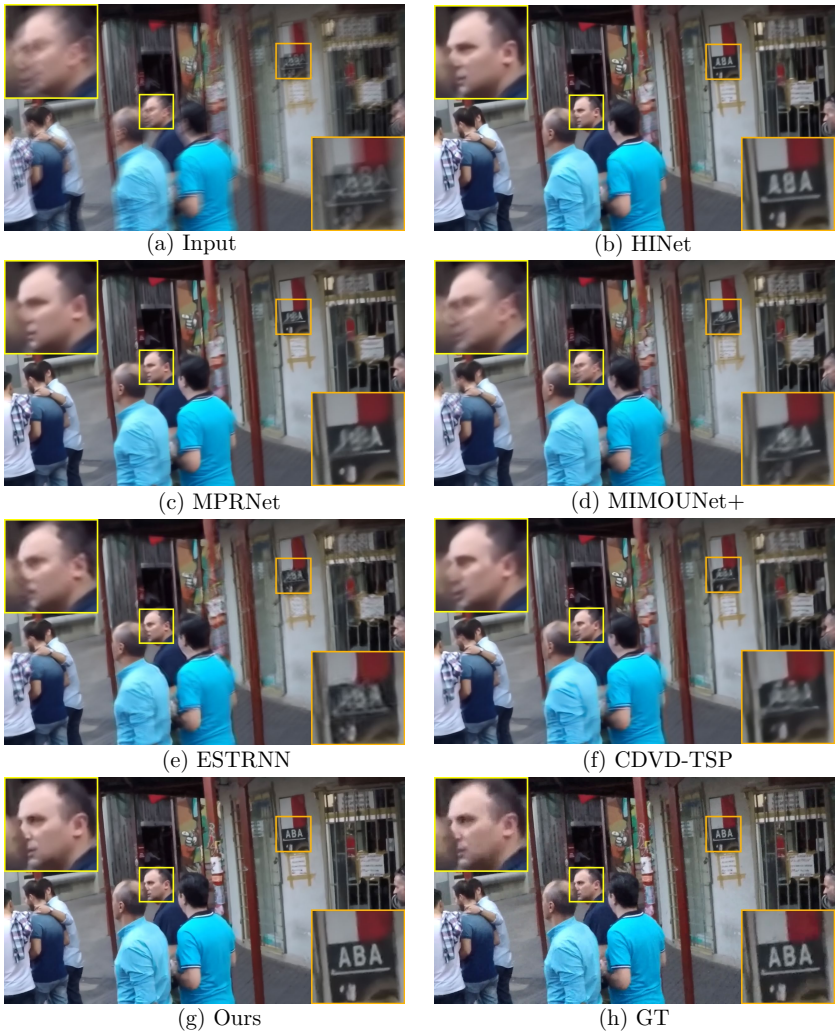


Fig. 10: Visual comparison of unknown exposure time blurry video frames on our real-world event datasets.



(a) Input (b) HINet (c) MPRNet (d) MIMOUNet+ (e) ESTRNN (f) CDVD-TSP (g) Ours (h) GT

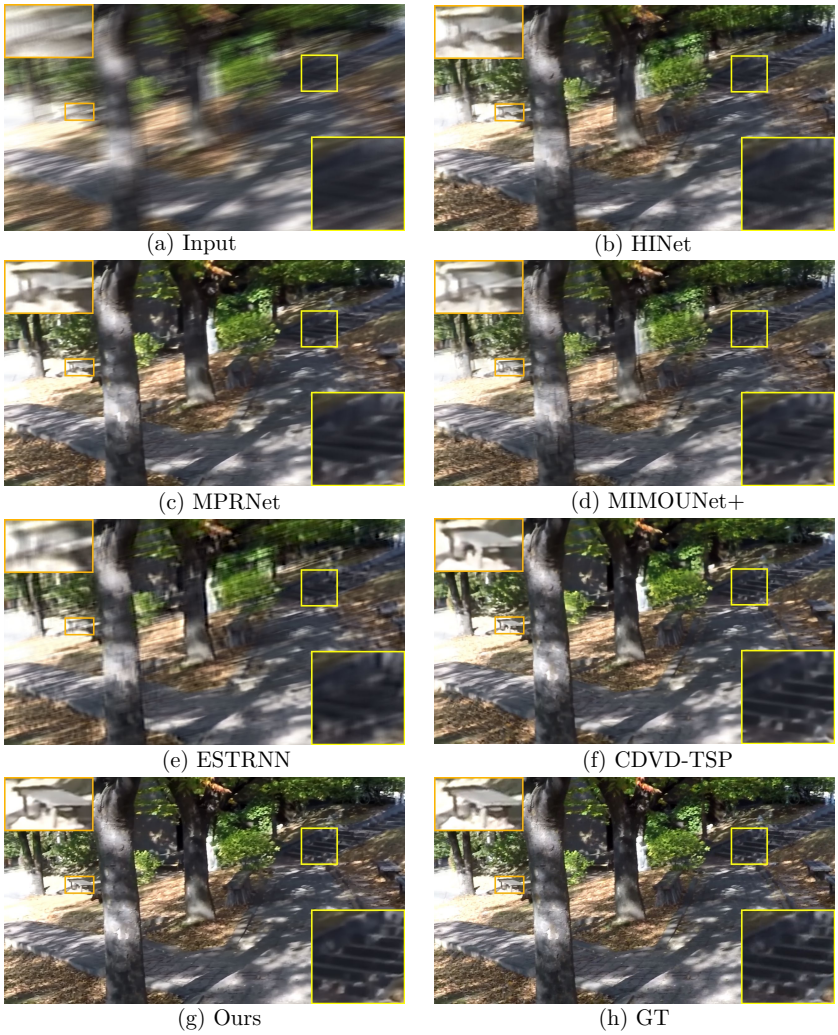
Fig. 11: Visual comparison on the GoPro-15fps datasets.

(a) Input (b) HINet (c) MPRNet (d) MIMOUNet+ (e) ESTRNN (f) CDVD-TSP (g) Ours (h) GT

Fig. 12: Visual comparison on the GoPro-15fps datasets.

(a) Input (b) HINet (c) MPRNet (d) MIMOUNet+ (e) ESTRNN (f) CDVD-TSP (g) Ours (h) GT

Fig. 13: Visual comparison on the GoPro-15fps datasets.



(a) Input (b) HINet (c) MPRNet (d) MIMOUNet+ (e) ESTRNN (f) CDVD-TSP (g) Ours (h) GT

Fig. 14: Visual comparison on the GoPro-15fps datasets.

References

1. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits **49**(10), 2333–2341 (2014)
2. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 182–192 (2021)
3. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
4. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. arXiv preprint arXiv:2108.05054 (2021)
5. Gehrig, D., Rüegg, M., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. IEEE Robotics and Automation Letters **6**(2), 2822–2829 (2021)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
7. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3320–3329 (2020)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
10. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.S.: Learning event-driven video deblurring and interpolation. In: ECCV (8). pp. 695–710 (2020)
11. Mostafavi, M., Choi, J., Yoon, K.J.: Learning to super resolve intensity images from events. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE/CVF (2020)
12. Mostafavi, M., Wang, L., Yoon, K.J.: Learning to reconstruct hdr images from events, with applications to depth and flow prediction. International Journal of Computer Vision **129**(4), 900–920 (2021)
13. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017)
14. Pan, J., Bai, H., Tang, J.: Cascaded deep video deblurring using temporal sharpness prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3043–3051 (2020)
15. Pan, J., Bai, H., Tang, J.: Cascaded deep video deblurring using temporal sharpness prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
16. Paredes-Vallés, F., de Croon, G.C.: Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. arXiv preprint arXiv:2009.08283 (2020)

17. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. In: British Machine Vision Conference (BMVC). British Machine Vision Association (BMVA) (2018)
18. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3857–3866 (2019)
19. Shang, W., Ren, D., Zou, D., Ren, J.S., Luo, P., Zuo, W.: Bringing events into video deblurring with non-consecutively blurry frames. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4531–4540 (October 2021)
20. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Kleeman, L., Mahony, R.: Reducing the sim-to-real gap for event cameras. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. pp. 534–549. Springer (2020)
21. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
22. Wang, L., , S.M.M.I., Ho, Y.S., Yoon, K.J.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
23. Wang, L., Chae, Y., Yoon, K.J.: Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In: ICCV (2021)
24. Wang, L., Chae, Y., Yoon, S.H., Kim, T.K., Yoon, K.J.: Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 608–619 (2021)
25. Wang, L., Kim, T.K., Yoon, K.J.: Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In: CVPR. pp. 8315–8325 (2020)
26. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
27. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. arXiv preprint arXiv:2102.02808 (2021)
28. Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5978–5986 (2019)
29. Zhang, S., Zhang, Y., Jiang, Z., Zou, D., Ren, J., Zhou, B.: Learning to see in the dark with events. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 666–682. Springer (2020)
30. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B.: Efficient spatio-temporal recurrent neural network for video deblurring. In: European Conference on Computer Vision. pp. 191–207. Springer (2020)
31. Zou, Y., Zheng, Y., Takatani, T., Fu, Y.: Learning to reconstruct high speed and high dynamic range videos from events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2024–2033 (2021)