Event-Based Fusion for Motion Deblurring with Cross-modal Attention
Supplementary Material

Lei Sun[1,2] Christos Sakaridis[2] Jingyun Liang[2] Qi Jiang[1] Kailun Yang[3]
Peng Sun[1] Yaozu Ye[1] Kaiwei Wang[1] Luc Van Gool[2,4]

[1]Zhejiang University [2]ETH Zürich [3]KIT [4]KU Leuven

This document provides additional materials to supplement our main manuscript. We first present more details about EFNet in Sec. 1, and then we show more information about the REBlur dataset in Sec. 2. In Sec. 4, we present more results on the GoPro [4] and REBlur datasets. Finally, we discuss potential negative impacts in Sec. 5.

1 More Details on EFNet

1.1 Supervision Attention Module

To help the second stage of EFNet also gain access to the input image, we place a supervision attention module [10] between the two stages of EFNet. As illustrated in Fig. 1, the feature maps from the first stage $\mathbf{F}_{\text{in}} \in \mathbb{R}^{H\times W\times C}$ first generate a residual image, via a 1×1 convolution. H, W, and C are the height, width, and number of channels. The blurry image is added with the residual image and we obtain the predicted deblurred image from the first stage. Then, after a 1×1 convolution and sigmoid function, the resulting attention maps are added to the identity mapping path. Finally, the output $\mathbf{F}_{\text{out}} \in \mathbb{R}^{H\times W\times C}$ is concatenated with the feature maps in the second stage.

1.2 Loss Function

We adopt the Peak Signal-to-Noise Ratio (PSNR) loss as the loss function for training our network. The definition of PSNR is:

$$PSNR(x,y) = 20log_{10}\frac{MAX}{\sqrt{MSE}}, \tag{1}$$

where MAX is the maximum possible pixel value of the image, and MSE is the mean squared error:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[x(i,j)-y(i,j)]^2, \tag{2}$$

where m, n are the number of rows and the number of columns in the images x and y, respectively.

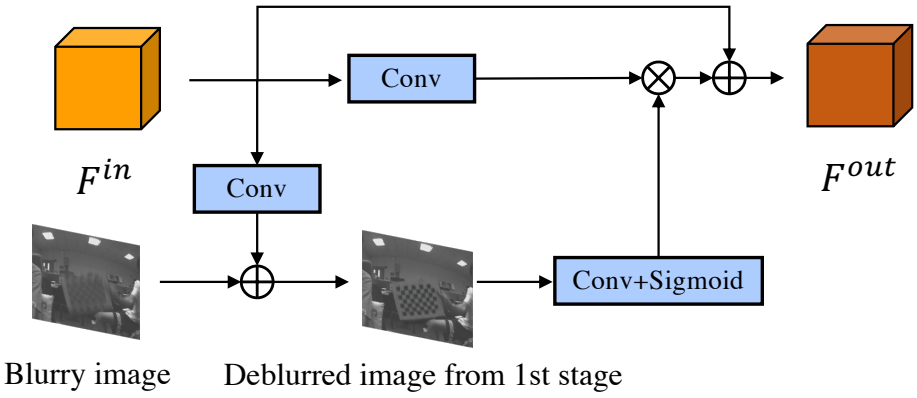


Fig. 1: The architecture of the supervision attention module [10].

The definition of PSNR loss in our model is:

$$Loss = -\frac{1}{2}\sum_{i=1}^{2} PSNR(X_i, Y) \tag{3}$$

where X_i is the predicted image of the i-th stage in EFNet, and Y is the ground-truth image.

2 More Details on REBlur dataset

2.1 Dataset Distribution

To enhance the generalization of the network for different objects and moving processes, REBlur includes 12 kinds of linear and nonlinear motions for 3 different moving patterns and for the camera itself, as detailed in Fig. 2. The dataset consists of 36 sequences and 1469 groups of blurry-sharp image pairs with associated events, where 486 pairs are used for training and 983 for testing. We include an additional set of 4 sequences including extreme blur, without ground truth. Please refer to the supplement for more details on REBlur.

2.2 Event Camera Detail

We choose Insightness Seem 1 camera, which is a Dynamic and Active Pixel Vision Sensor (DAVIS) outputting time-aligned 360×262 gray images and event streams. For the convenience of data processing, we discard the last two columns, so the image size is 360×260 in the dataset. Table 1 shows detailed information about our camera.

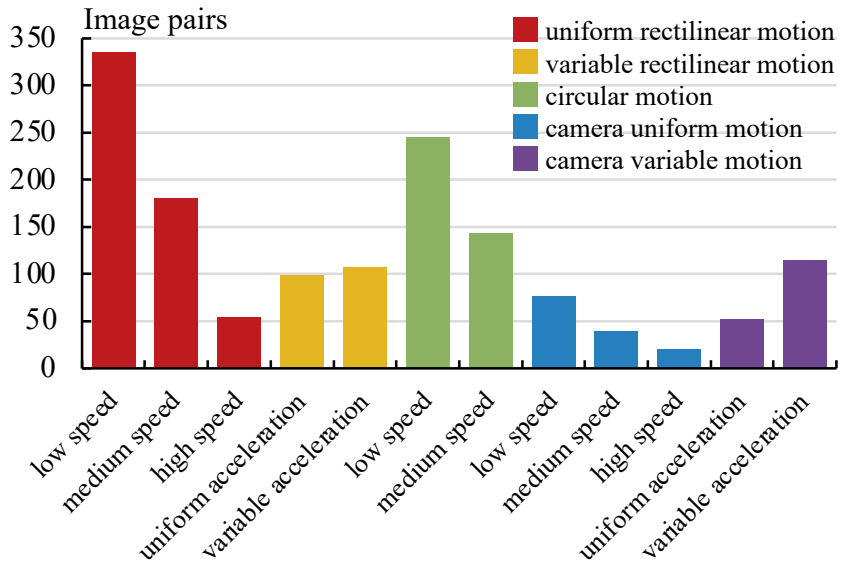


Fig. 2: Distribution of different motion categories in our REBlur dataset.

Table 1: **Detailed information** about Insightness Seem 1 camera.

Generation	**SEEM1**	
Lens	F2.0, 2.5mm, FOV 130°	
Spatial resolution	320 (horizontal) × 262 (vertical) active pixels	
Optical format	1/3.2 inch, pixel field is centered	
Pixel pitch	13um	
Data type	**Change detection events**	**Image frames**
Temporal resolution	Up to 10kHz (configurable)	Up to 30Hz (configurable)
Dynamic range	>98 dB	50 dB
Shutter mode	-	Electronic global shutter
Sensitivity/ Intensity resolution	Configurable threshold down to 50% contrast	10-bit on-chip ADCs
Latency	<1 ms for >150 lux scene illumination <5 ms for >30 lux scene illumination	-
Readout	USB 2.0 and MIPI-CSI 2 (under development)	
Readout bandwith	∼12 Meps (USB 2.0) >40 Meps (MIPI)	

2.3 Data Capture

We adopt the two-shot strategy in blurry-sharp image collection. To ensure the motion of the object or the motion of the camera being the only factor that changes in the two shots, we keep the illumination still and other objects in the field of view of the camera still.

In the first shot, we capture images with motion blur for the pattern on the slide-rail and corresponding event streams. In the second shot, according to the timestamp t_s of the blurry images, we select events within the time range $[t_s - 125\mu s, t_s + 125\mu s]$ and visualize these events in the preview of the sharp image capture program. Referring to the edge information from high-temporal-resolution events, we can relocate the slide-rail to the coordinate corresponding to the timestamp t_s by an electronic-controlled stepping motor and then capture the latent sharp image.

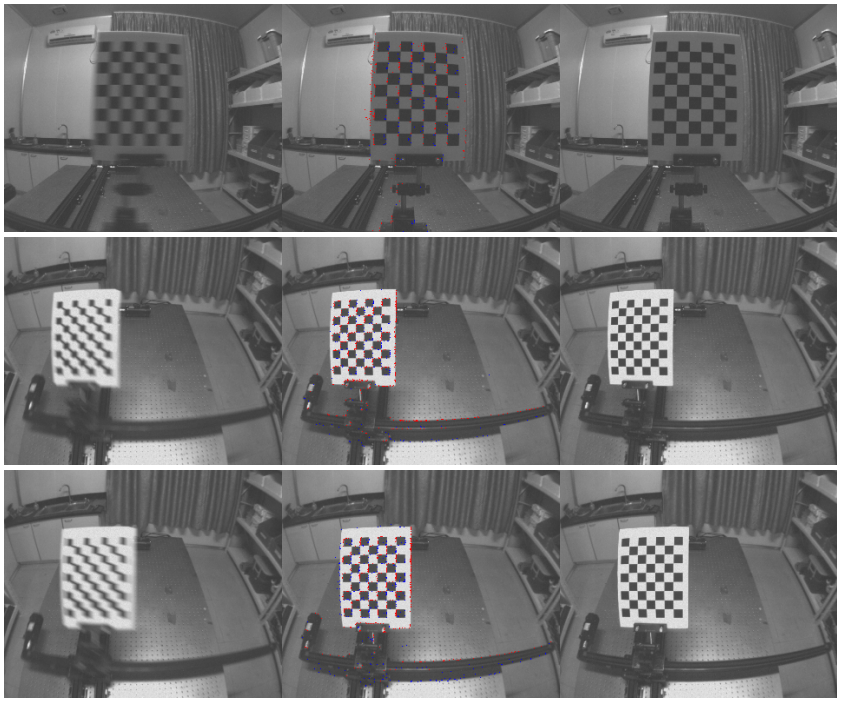
Fig. 3 shows blurry images and corresponding sharp ground truth images with visualized events in the time range $[t_s - 125\mu s, t_s + 125\mu s]$. In all sequences, the object movement distance in the direction perpendicular to the optical axis corresponds to a movement distance of less than one pixel on the sensor in the image space.

2.4 Dataset evaluation and analysis

Rationality: The rationality and accuracy of our two-shot capturing method can be supported by the following two facts: (1) **Latent sharp image alignment.** The positioning error of the high-precision slide-rail (0.05mm) and the maximum displacement of the pattern within the exposure time corresponds to projected distances of less than one pixel on the sensor in the image plane. (2) **Background consistency.** Due to the highly stable optical laboratory environment, the illumination is strictly constant during the alignment process and the acquisition of blurry images. For the background part of blurry and sharp images in the same pair, PSNR is 44.94, indicating that they are highly consistent.
Relative PSNR/SSIM results: REBlur contains both object- and camera-motion blur. For object-motion blur, the background of the blurry image is identical to that of the sharp one, which results in overall higher scores for all methods on REBlur than on GoPro, as they simply need to leave the background unchanged. Thus, the average PSNR/SSIM value is higher than that in REBlur, but it is harder to improve scores on REBlur than on GoPro.

3 More Training details

Dataset split: We use 2103 pairs for training and 1111 pairs for testing for GoPro [4] (standard setting). We train EFNet on GoPro dataset with 4 Nvidia Titan RTX GPUs and last for about 40 hours. We arrange 486 pairs for training and 983 for testing for REBlur. The training and test set each contain three different moving speeds, and distinct moving objects. Fine-tuning on REBlur is conduct on 1 Nvidia Titan RTX GPU and last for about 10 minutes.



(a) Blurry Image (b) Visual Event (c) Ground Truth

Fig. 3: **Example images from the dataset collection procedure for REBlur.** (a): Blurry images in the first shot. (b): In the second shot, we align the pattern with the visualized events. (c): The captured ground truth.

Training strategies: Fine-tuning on REBlur uses the same training strategies as on GoPro, except for the l.r. and iterations. For the other image-only and event-based methods, all the training and fine-tuning settings are the same for a fair comparison.

Synthetic events in GoPro. We use ESIM [6] to produce synthetic events from sharp image sequences. This is an common practice for the other event-based image deblur methods [2,3,7,5].

4 More Qualitative Results

In this section we compare our EFNet with SRN [8], HINet [1], MPRNet [10], SRN+ (the enhanced version of SRN), HINet+ (the enhanced version of HINet), and BHA [5]. We show more qualitative results in Fig. 4 and Fig. 4.

4.1 GoPro Dataset

As illustrated in Fig. 4, our method restores the structural elements (*i.e.* numbers, characters, and lines) and detailed textures (*i.e.* face detail). Compared

to state-of-the-art image-based and event-based methods, our method achieves better restoration results.

4.2 REBlur Dataset

Fig. 4 shows more qualitative results from the REBlur dataset. Image-only methods (SRN [8], HINet [1], MPRNet [10]) perform poorly in such severe blurry conditions. Event-based methods are more robust, but BHA [5] is prone to noise. SRN+ and HINet+ show artifacts because of the insufficient use of event information. Our EFNet achieves the best performance.

5 Potential Negative Societal Impacts

Since event cameras will likely go to mass production, some of the cell phones may be equipped with this advanced sensor in the near future and our event-based deblurring algorithm may be applied in these cell phones. Our algorithm improves image deblurring performance compared to image-only methods, especially in severe blurry conditions. After mitigating motion blur in the images, one potential negative impact is that intrusive shots are made easier and thus cause bad social effects. This can be alleviated by forcing shutter sound and other methods.

6 Limitations

Although we have achieved impressive deblurring results in most situations, EFNet also shows performance degradation in the most adverse blurring conditions. Here we show some failure examples on the additional set of the REBlur dataset in Fig. 5 and Fig. 4 (b).

These are the most severe blur conditions. Even the basic shape of the moving objects cannot be distinguished based only on the blurry image. Although the deblurring performance is not ideal, the basic pattern and shape of the moving object can be recognized. The main reason for failure in these examples is due to the limitation of the hardware: (1) **The temporal resolution of event camera is not enough for such blurry conditions.** This is determined by the event camera itself and the degradation would be alleviated if the hardware improved. (2) **The refractory period makes each pixel not able to react to next intensity changes after last events in a short time,** and this is also a random factor [9]. Because we train models on the synthetic events from ESIM [6], this may be mitigated if more simulation settings are added in ESIM [6].

References

1. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: HINet: Half instance normalization network for image restoration. In: CVPRW (2021) 5, 6, 7, 8, 9, 10

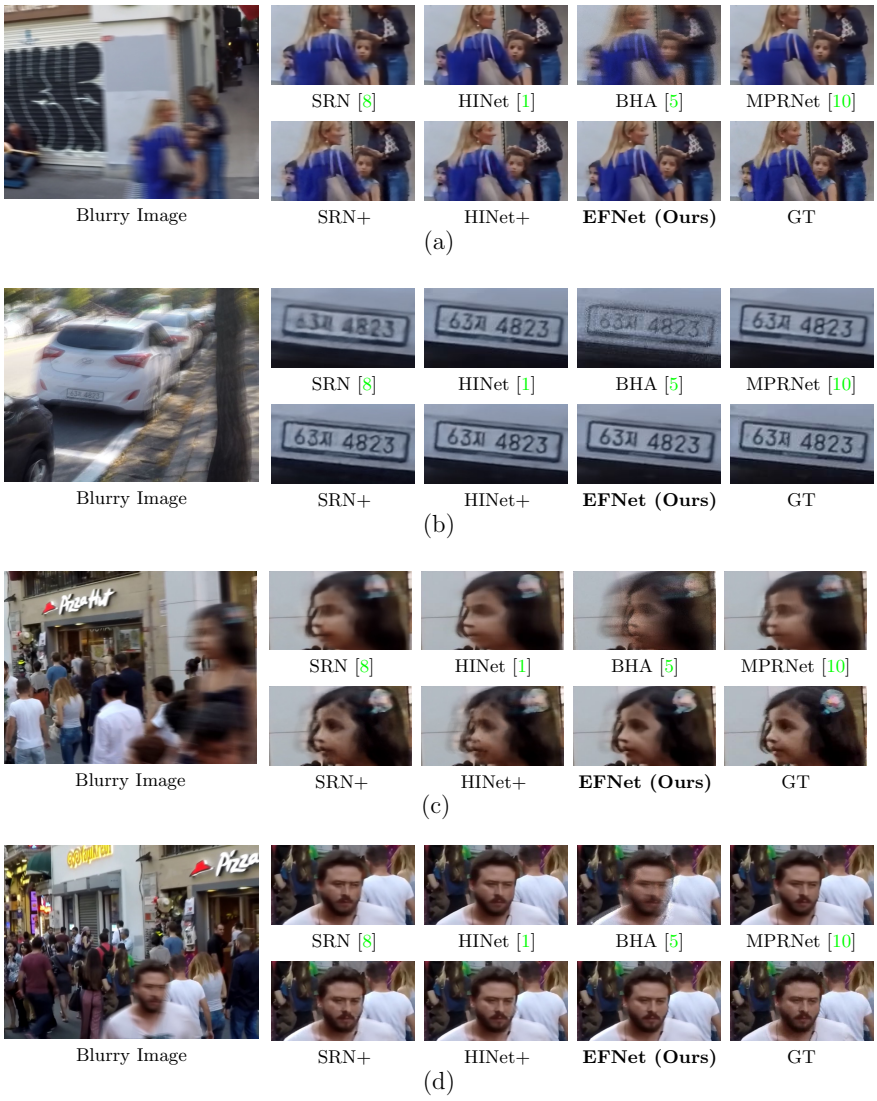


Fig. 4: **More visual comparisons on the GoPro dataset.** SRN+ and HINet+: event enhanced versions of SRN [8] and HINet [1] respectively. Our method restores the structural elements (*i.e.* numbers, characters, and lines) and detailed textures (*i.e.* face detail). Compared to state-of-the-art image-based and event-based methods, our method achieves better restoration results. Best viewed on a screen and zoomed in.

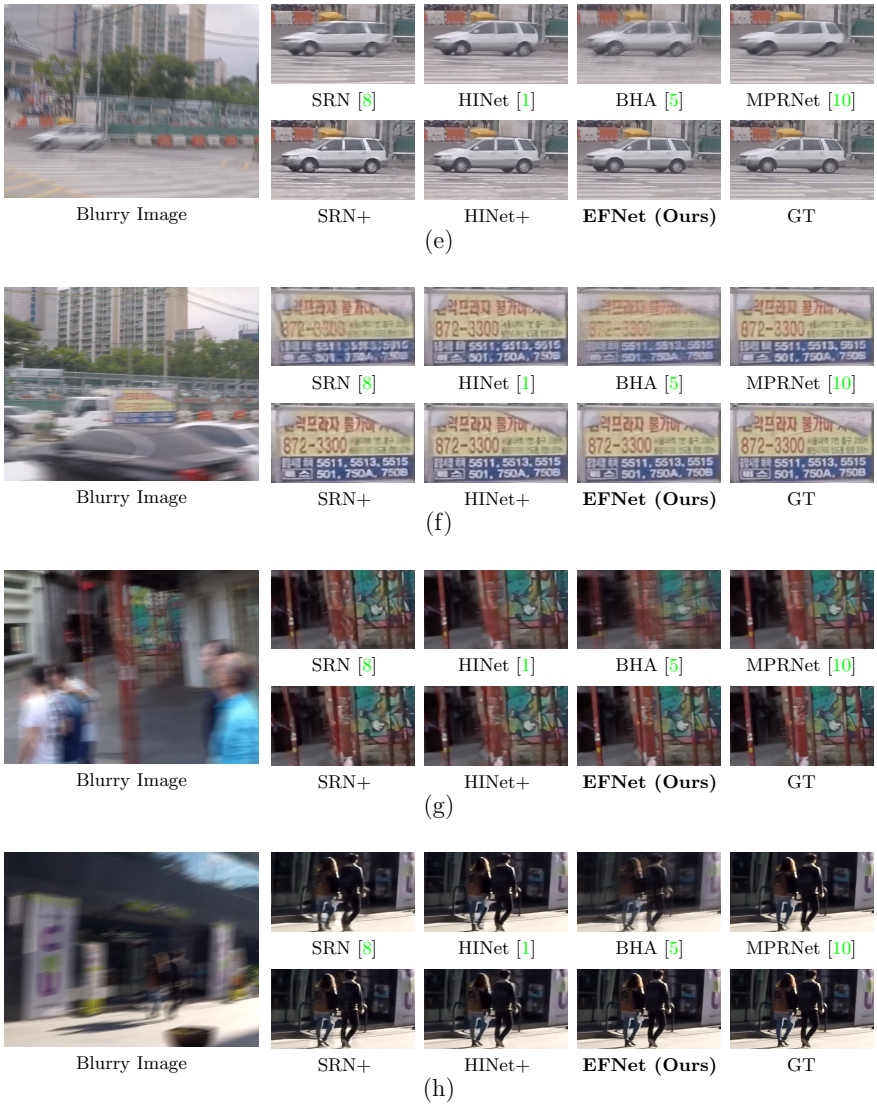


Fig. 3: (continued) **More visual comparisons on the GoPro dataset.** SRN+ and HINet+: event enhanced versions of SRN [8] and HINet [1] respectively. Our method restores the structural elements (*i.e.* numbers, characters and lines) and detailed textures (*i.e.* face detail). Compared to image-based and event-based state-of-the-art methods, our method achieves better restoration results. Best viewed on a screen and zoomed in.

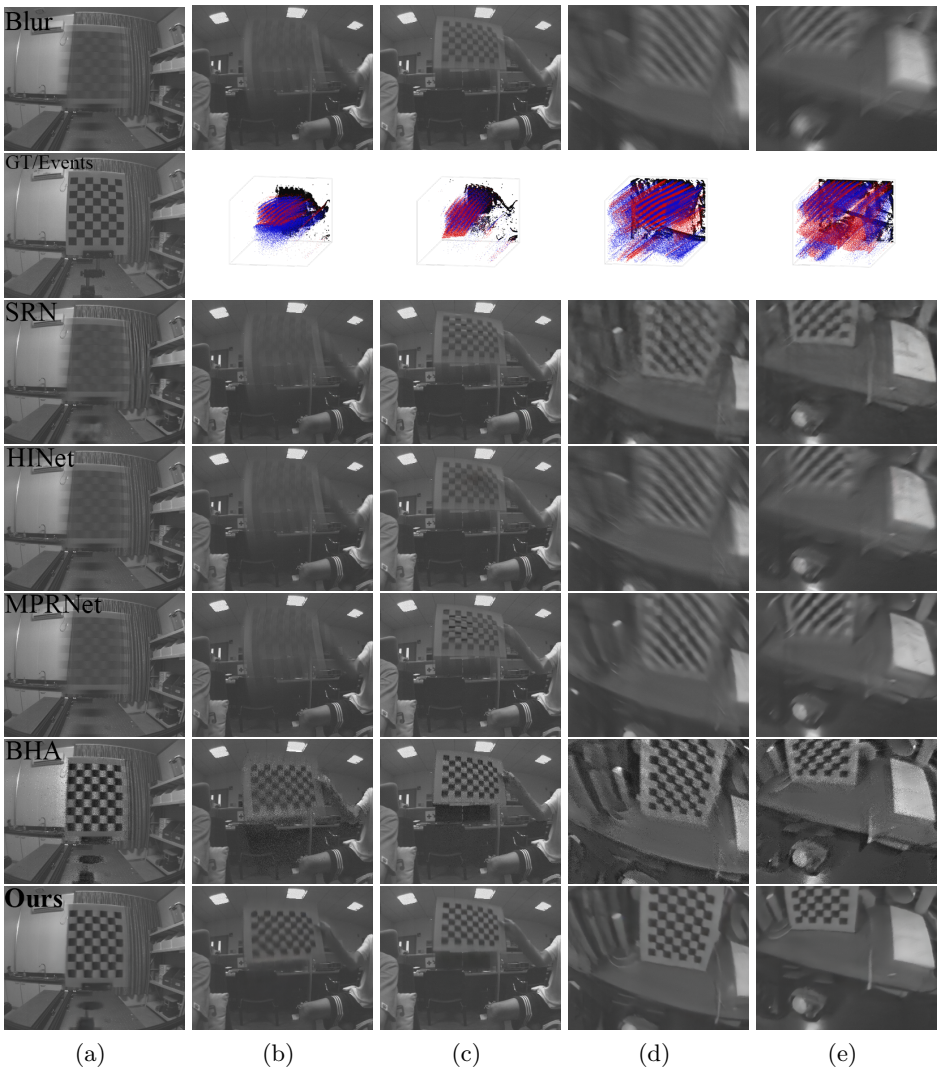
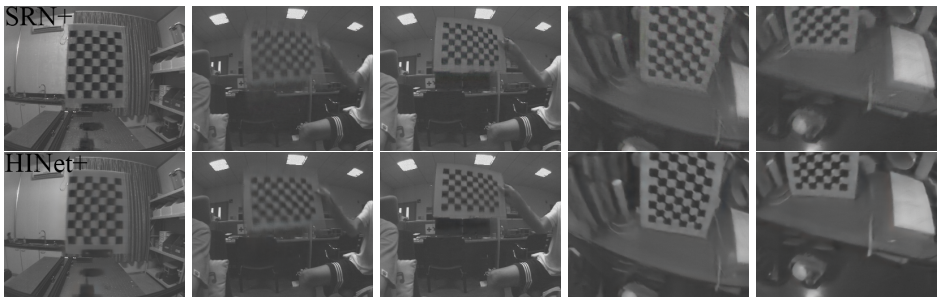


Fig. 4: **More visual comparisons on the REBlur dataset.** SRN+ and HINet+: event enhanced versions of SRN [8] and HINet [1] respectively. (a) is from the test set and (b), (c), (d), and (e) are from the additional set of REBlur. Image-only methods (SRN [8], HINet [1], MPRNet [10]) perform poorly in such severe blurry conditions. Event-based methods are more robust, and our EFNet achieves the best performance. Best viewed on a screen and zoomed in.



(a) (b) (c) (d) (e)

Fig. 4: (continued) **More visual comparisons on the REBlur dataset.** SRN+ and HINet+: event enhanced versions of SRN [8] and HINet [1] respectively. (a) is from the test set and (b), (c), (d), and (e) are from the additional set of REBlur. Image-only methods (SRN [8], HINet [1], MPRNet [10]) perform poorly in such severe blurry conditions. Event-based methods are more robust, and our EFNet achieves the best performance. Best viewed on a screen and zoomed in.

(a) Blurry Image (b) Deblurred Image

Fig. 5: **Failure example in the most severe blur condition.** In the blurry image, the chessboard cannot be delineated. The deblurred image shows some artifacts, and some black squared of the chessboard are merged.

2. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: CVPR (2020) 5
3. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.: Learning event-driven video deblurring and interpolation. In: ECCV (2020) 5
4. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017) 1, 4
5. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: CVPR (2019) 5, 6, 7, 8
6. Rebecq, H., Gehrig, D., Scaramuzza, D.: ESIM: an open event camera simulator. In: CoLR (2018) 5, 6

7. Shang, W., Ren, D., Zou, D., Ren, J.S., Luo, P., Zuo, W.: Bringing events into video deblurring with non-consecutively blurry frames. In: ICCV (2021) 5
8. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: CVPR (2018) 5, 6, 7, 8, 9, 10
9. Xu, F., Yu, L., Wang, B., Yang, W., Xia, G.S., Jia, X., Qiao, Z., Liu, J.: Motion deblurring with real events. In: ICCV (2021) 6
10. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021) 1, 2, 5, 6, 7, 8, 9, 10