Event-Based Frame Interpolation with Ad-hoc Deblurring
Supplementary Material

Lei Sun[1,2] Christos Sakaridis[2] Jingyun Liang[2] Peng Sun[1] Jiezhang Cao[2] Kai Zhang[2]
Qi Jiang[1] Kaiwei Wang[1] Luc Van Gool[2,3]

[1]State Key Laboratory of Modern Optical Instrumentation, Zhejiang University
[2]Computer Vision Lab, ETH Zürich
[3]PSI, KU Leuven

A. More Details on REFID

Fig. 7 shows the detailed architecture of the proposed Event Recurrent (EVR) block. The feature maps from the previous blocks are concatenated with the previous state, and then sent to the residual blocks. In the end, after the convolutional layer and the activation layer, the new state is propagated to the next recurrent iteration. The feature maps are down-sampled to half of the original size, and sent to the next block.

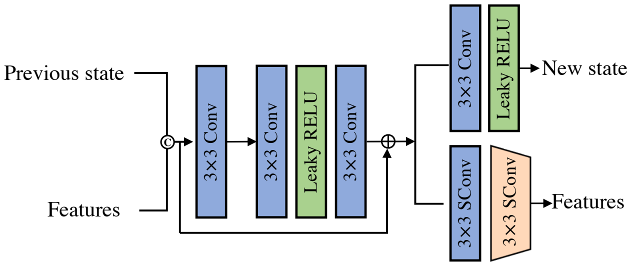

Figure 7. Detailed architecture of the proposed Event Recurrent (EVR) block.

B. More Training Details

B.1. Loss Function

We use the Charbonnier loss [3], which is a widely used loss function in low-level vision tasks.

$$L_{\text{Charbonnier}} = \sqrt{(y-\hat{y})^2+\epsilon^2}, \tag{1}$$

where y and $\hat{y}$ are the ground-truth frame and predicted frame, respectively. ϵ is set to 10^{-6} in our experiments.

Compared with L_1 and L_2 losses, the Charbonnier loss is more robust, which better handles outliers [3].

C. Testing on Different Skips

All the experimental settings for testing are kept the same as in training in the main paper. In this section, we discuss the performance when the testing setting is different from the training setting, with respect to the number of the interpolated frames (Sec. C.1, Sec. C.2) and the sharpness of the input videos (Sec. C.3).

C.1. Blurry Frame Interpolation

In Tab. 5, different rows correspond to different training-time skips, while different columns correspond to different test-time skips. The REFID trained on blurry frames with 3 skips achieves an SSIM of 0.972 in the 1-skip setting, which is only marginally lower than the model trained with 1 skipped frame (0.973). This finding extends to the reverse situation, in which REFID is trained on 1 skipped frame and tested on 3 skipped frames, which shows the robustness of our REFID to different testing settings.

Table 5. **Testing on different blurry frame skips**. Training and testing is performed on the GoPro [4] dataset.

Training settings	**3 skip**		**1 skip**	
	PSNR	**SSIM**	**PSNR**	**SSIM**
Trained on 3skip	35.47	0.971	34.12	0.972
Trained on 1skip	34.01	0.970	35.90	0.973

C.2. Sharp Frame Interpolation

In the sharp frame interpolation setting, for the models trained on 15 skipped frames, we also test their performance on 7, 3, and 1 skipped frame(s) in Tab. 6. Our REFID trained on sharp frames with 15 skips achieves PSNRs of 33.80/31.63/30.17 for 7/3/1 skipped frames, i.e., its performance decreases gracefully, which shows the generalization of our model to different testing settings.

Table 6. **Test on different sharp frame skips**. The model is trained on sharp 15-skip GoPro dataset.

Skipped frames	**15**	**7**	**3**	**1**
PSNR	35.63	33.80	31.63	30.17
SSIM	0.974	0.969	0.956	0.952

C.3. Training on Blurry Frame Interpolation and Testing on Sharp Frame Interpolation

The model architecture of the proposed REFID for blurry and sharp frame interpolation is identical. We show the result for 1 skipped frame of the model trained on blurry frames when it is tested on sharp frames in Tab. 7. Although the result shows some degradation compared to the model trained on sharp frames, performance is still competitive for sharp frame interpolation.

Table 7. **Evaluation on different blur conditions.** Test on sharp frame skips. Models are trained on the blurry HighREV dataset with 1 skipped frame and tested on the sharp HighREV dataset with 1 skipped frame.

Methods	RIFE [2]	**REFID (Ours)**
PSNR	28.74	**31.16**
SSIM	0.837	**0.929**

D. More Details on HighREV Dataset

Data split. In [5], to construct an event-based high-resolution dataset, the authors combine a synchronized, high-resolution (1280 × 720) event camera with an RGB camera to make a hybrid sensor. However, the alignment of the two sensors introduces error both in the temporal axis and the spatial axis. Our HighREV dataset is collected using one sensor that outputs both events and RGB frames at the same time, with a resolution of 1632 × 1224. Because it is a Dynamic and Active VIsion Sensor (DAVIS) [1], events and RGB images are aligned by design.

Fig. 8 shows the proportion of the train and test sets of HighREV, and the proportions of indoor and outdoor scenes in the two sets. We keep the ratio of indoor and outdoor scenes approximately the same in the train set and the test set. 70% of the video sequences are used for training and 30% for testing. In our experiments, the train set is used for fine-tuning models on HighREV.

Fig. 9 shows the detailed image distribution for each sequence. To preserve anonymity, we replace the name of the place or institute with "*".

For the collection of the dataset, the exposure time of the camera is set to 15ms and the f-stop of the lens is set to 2. The frame rate of the APS image is set to 25.

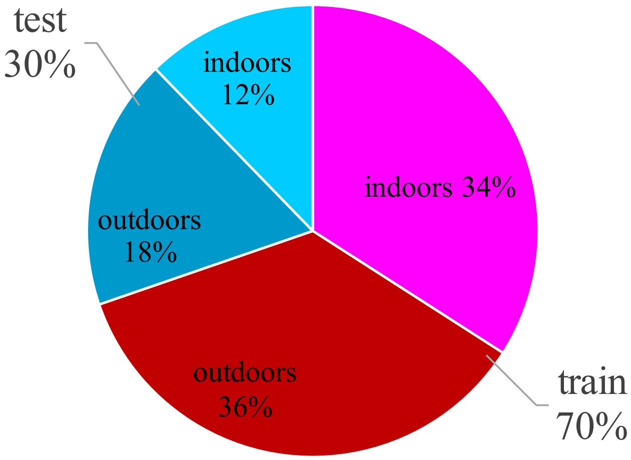

Figure 8. **Data split** of the HighREV dataset.

Event camera details. The resolution for RGB frames and event frames is 1632 × 1224. The size of the pixel pitch is $1.89\mu m \times 1.89\mu m$. The minimun illumination is lower than 0.1 lux. The Chief Ray Angle (CRA) for the lens is 34°.

E. Additional Qualitative Results

E.1. Blurry Frame Interpolation

In this section, we show more qualitative results of REFID on event-based blurry frame interpolation on both our HighREV dataset and the GoPro dataset in Fig. 10 and 11, respectively. REFID is able both to effectively remove the blur that is present in the input and accurately capture the motion between the left and right frames.

E.2. Sharp Frame Interpolation

In this section, we show more qualitative results of REFID on event-based sharp frame interpolation on our HighREV dataset in Fig. 12. We observe that REFID correctly captures the motion between the two input frames in interpolating the intermediate frames.

E.3. Single Image Deblurring

In this section, we show more qualitative results of REFID on event-based single image deblurring on the GoPro [4] dataset in Fig. 13 and 14. Despite the intense blur that is present in most of the examples of Fig. 13 and 14, REFID produces sharp results which are faithful to the ground-truth sharp image and contain minimal artifacts.

F. Potential Negative Societal Impacts

Since event cameras will likely go to mass production, some of the cell phones may be equipped with this advanced

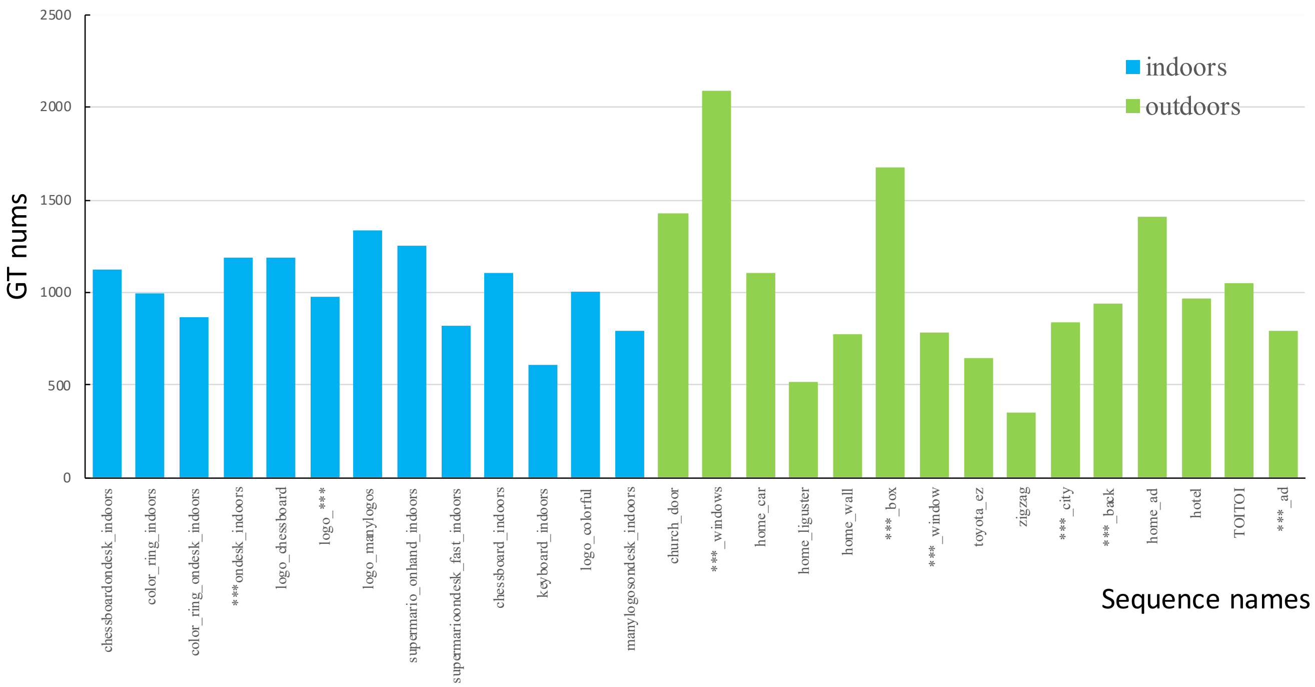

Figure 9. **Data distribution** of the HighREV dataset.

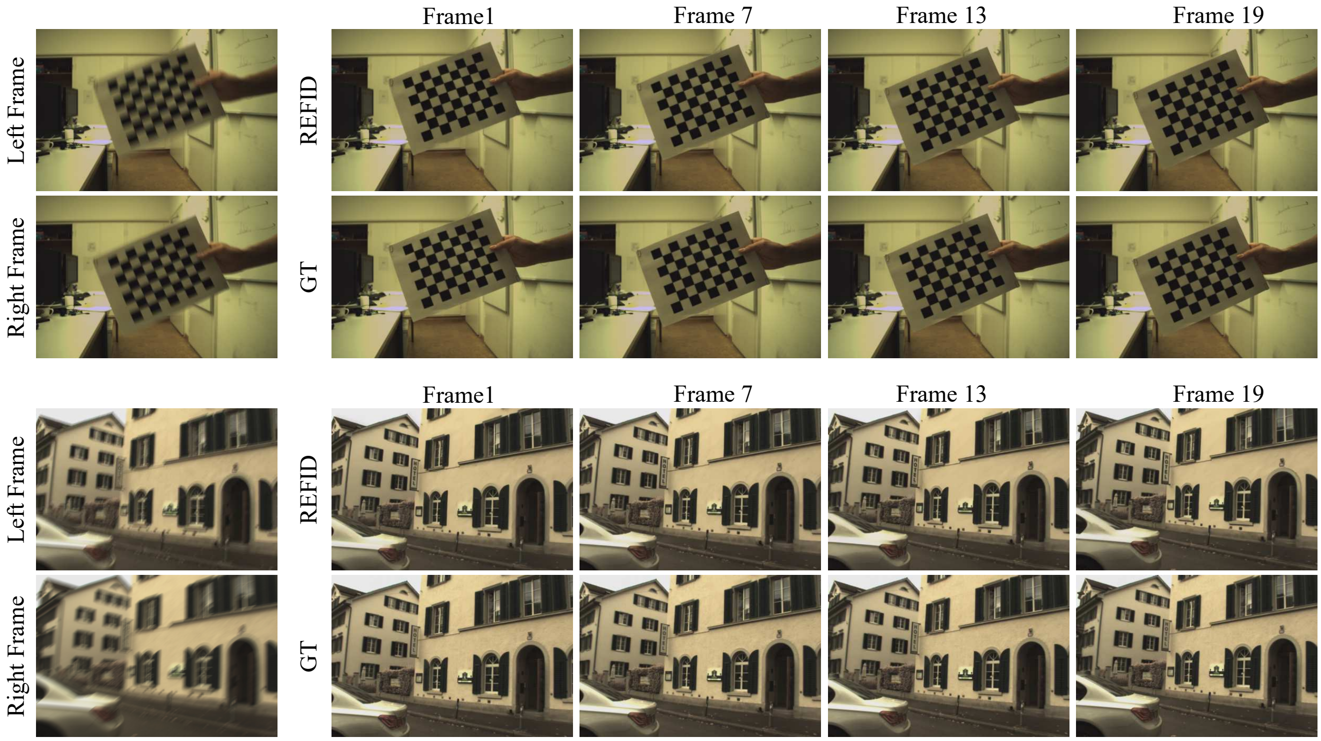

Figure 10. **Additional event-based blurry frame interpolation results** of our REFID on HighREV. We select the 1st, 7th, 13th, and 19th frames from the 25 total frames for visualization.

sensor in the near future and our event-based deblurring algorithm may be applied on these cell phones. Our algorithm

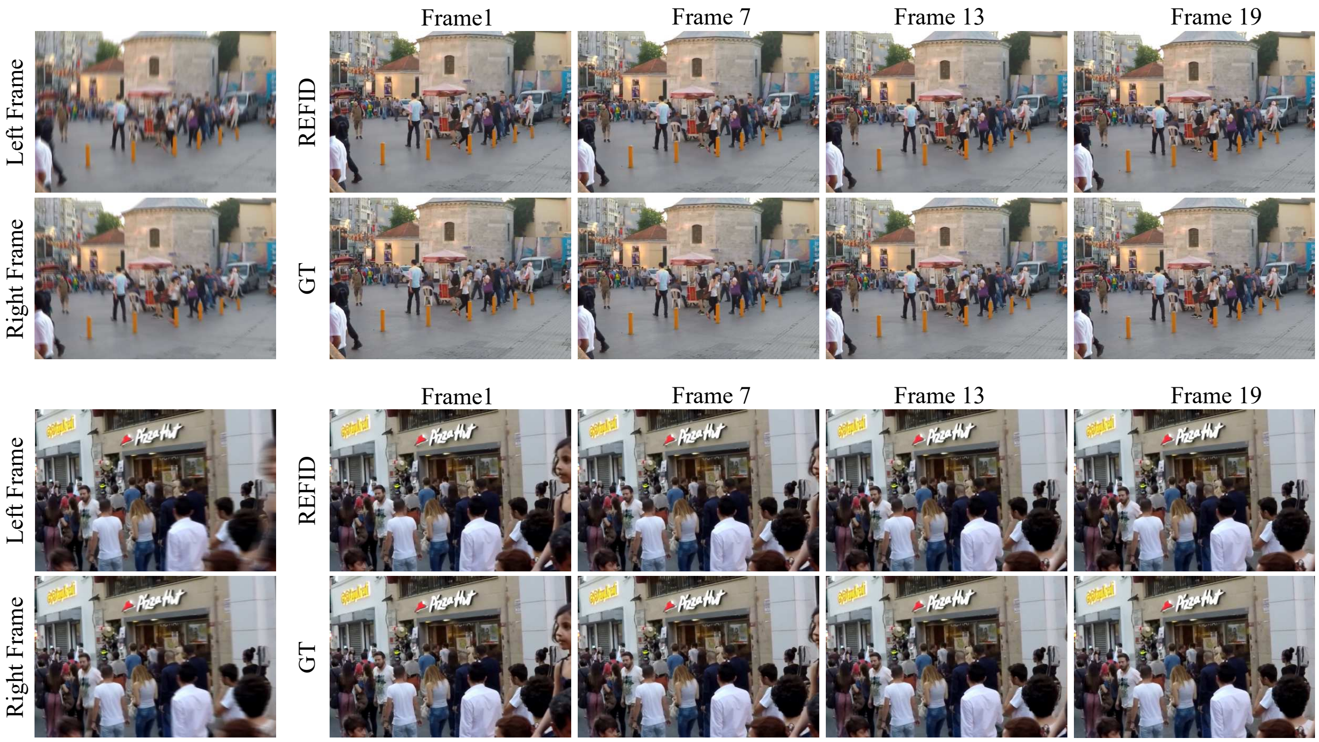

Figure 11. **Additional event-based blurry frame interpolation results** of our REFID on GoPro. We select the 1st, 7th, 13th, and 19th frames from the 25 total frames for visualization.

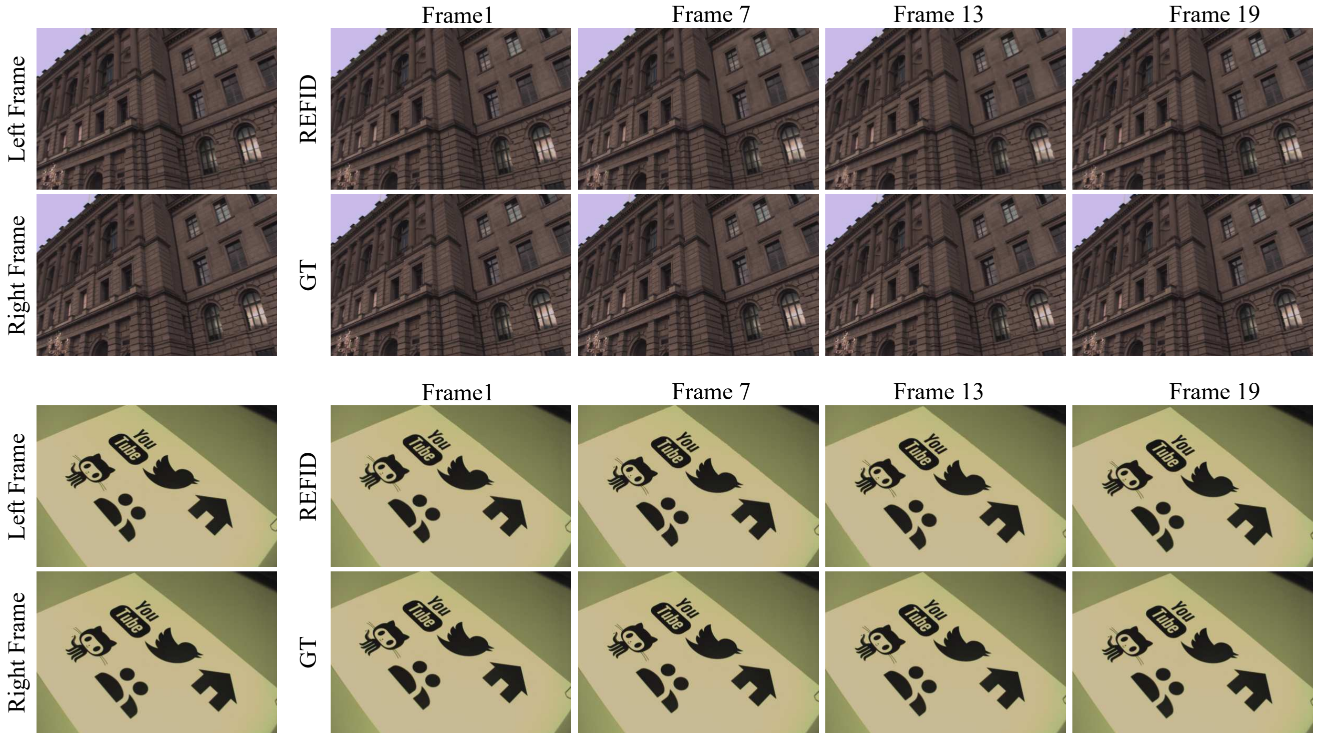

Figure 12. **Additional event-based sharp frame interpolation results** of our REFID on HighREV. We select the 1st, 5th, 9th, and 13th frames from the 25 total frames for visualization.

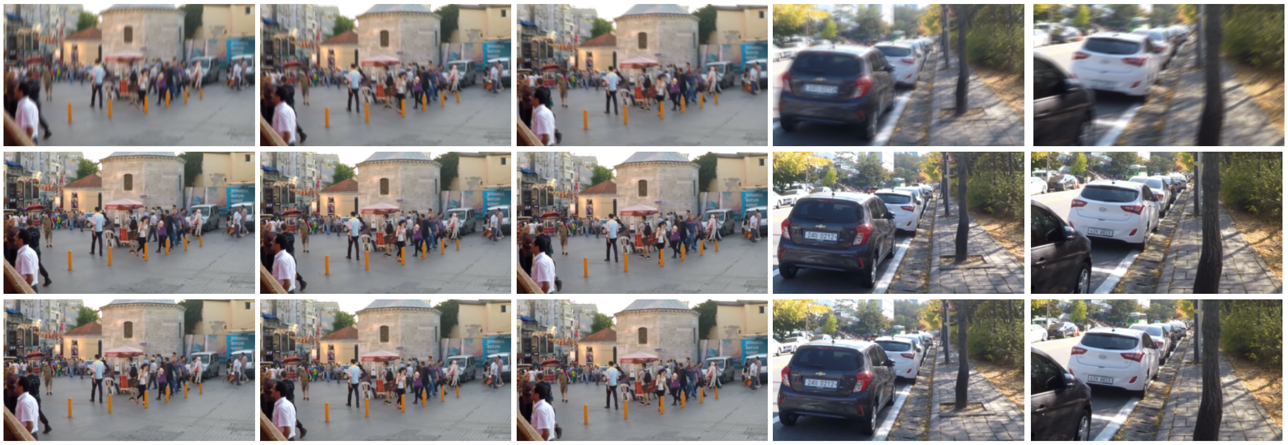

Figure 13. **Additional event-based single image deblurring results** of our REFID on the test set of GoPro. From top to bottom: blurry image, result of REFID, ground truth.

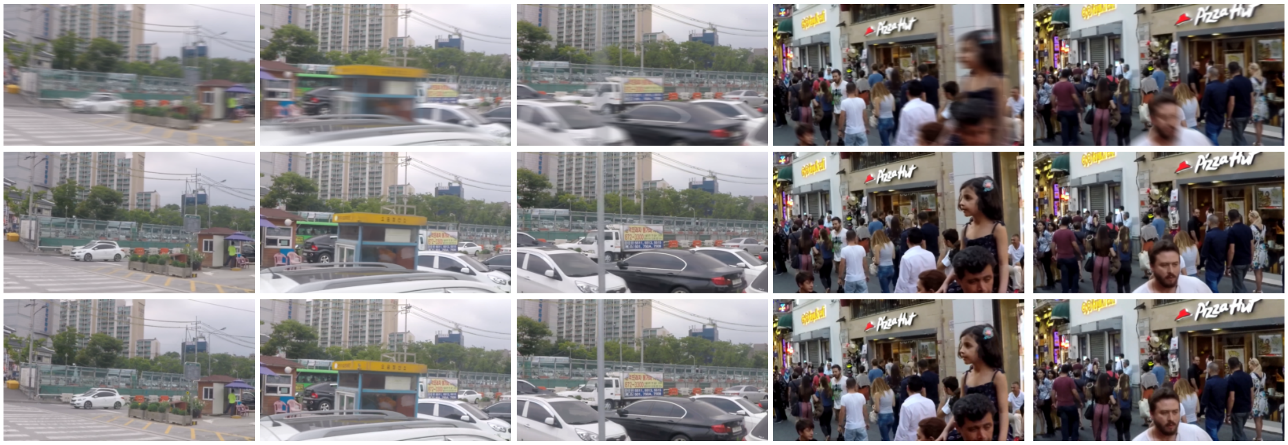

Figure 14. **Additional event-based single image deblurring results** of our REFID on the test set of GoPro. From top to bottom: blurry image, result of REFID, ground truth.

improves video interpolation and image deblurring performance compared to previous state-of-the-art methods, especially under severe blur. After mitigating motion blur in the images, one potential negative impact is that intrusive shots are made easier and thus cause bad social effects regarding privacy. This can be alleviated by forcing shutter sound or with other methods.

References

[1] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 dB 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 2014. 2

[2] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. RIFE: Real-time intermediate flow estimation for video frame interpolation. *arXiv preprint arXiv:2011.06294*, 2020. 2

[3] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018. 1

[4] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 2

[5] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16155–16164, 2021. 2