Supplementary Material for "EVS-assisted joint Deblurring, Rolling-Shutter Correction and Video Frame Interpolation through Sensor Inverse Modeling"

Rui Jiang*, Fangwen Tu*, Yixuan Long, Aabhaas Vaish, Bowen Zhou, Qinyi Wang
Wei Zhang, Yuntan Fang, Luis Eduardo Garcia Capel, Bo Mu, Tiejun Dai, Andreas Suess

OMNIVISION

*Equal contribution

Note that images displayed in the main paper have been enlarged for improved visibility. In the supplementary material, full-resolution images are presented. For video comparative results, please see https://docs.google.com/document/d/1LVuaZY847m4z0QmAGIC69liZVgcUZkDbt_jXCj6Pz7Y/edit.

7. Details of Datasets

In this section, we present the datasets used in this paper. The dataset details are summarized in Table 3.

Table 3. List of the datasets for evaluation. "RS" and "GS" denote "Rolling Shutter" and "Global Shutter", respectively; "CFA" stands for "Color Filter Array"; "GT" means "Ground Truth".

Scene	CIS setting	EVS CFA	GT	Description
Simulation data				
basketball	RS, 120 FPS	color	10 kFPS	Text and high frequency texture (ball skin) appearing and disappearing on a spinning basketball
checkerboard	RS, 120 FPS	color	10 kFPS	A free falling and rotating checkerboard with strong color contrast
slingshot egg	RS, 120 FPS	color	10 kFPS	A ultra-fast pellet flying toward an egg, large V_{FE} delay
running man	RS, 120 FPS	color	10 kFPS	A running man with slow speed, small V_{FE} delay
fan	RS, 120 FPS	color	10 kFPS	A rotating fan with fast speed rigid motion and small color contrast
Measurement data - natural environments				
mono drone	RS, 121.8 FPS	monochrome	N.A.	A still drone with rotating blades
color drone	RS, 121.8 FPS	color	N.A.	A still drone with rotating blades
table tennis	RS, 121.8 FPS	monochrome	N.A.	Text appearing and disappearing on a spinning table tennis ball
badminton	RS, 121.8 FPS	color	N.A.	An upward-bouncing shuttlecock after falling
Measurement data - controlled environments				
rotating disk	RS, 121.8 FPS	monochrome	N.A.	A rotating Siemens star under varying levels of illumination and speed
HS-ERGB dataset				
spinning disk	GS, 150 FPS	monochrome	N.A.	A plate undergoing out-of-plane rotation
spinning umbrella	GS, 153 FPS	monochrome	N.A.	An umbrella with a regular pattern undergoing in-plane rotation

7.1. Simulation Data

The simulation data is generated by a CIS-EVS hybrid sensor simulator [4] which models EVS latency from both pixel front-end and peripheral readout circuitry. The simulator has been calibrated to mimic real sensor characteristics. Figure 12A illustrates the input and output of the simulator. The simulator processes a series of input images, which stem from high-speed camera captures at a frame rate of 10 kFPS. These input frames are also used as the ground truth for algorithm evaluation. The simulator outputs events at a resolution of 540×960 as well as blurry 1080p CIS frames with RS effect. The simulator ensures that CIS and EVS data are synchronized temporally, in correspondence to the operation of a hybrid CIS-EVS sensor. The CIS RS row exposure time is set to 1 ms. The EVS contrast threshold and refractory time are configured to 20 % and 20 µs, respectively.

7.2. Measurement Data

We use a hybrid CIS-EVS sensor [3] for data collection. Since we lack ground truth for measured video capture, only qualitative evaluation results are presented. Scenes from **natural environments** cover both indoor and outdoor sequences, as depicted in Figure 13A. The positive and negative triggering thresholds are configured to $23\,\%$ and $16\,\%$, respectively.

In order to achieve a **controlled environment**, we establish an experimental setup with two adjustable LED light sources and a Siemens star pattern positioned on the motorized plate. We proposed five different ambient lighting conditions ($50\,\mathrm{lx}$, $100\,\mathrm{lx}$, $300\,\mathrm{lx}$, $500\,\mathrm{lx}$, and $1000\,\mathrm{lx}$) and five rotational speeds ($76\,\mathrm{rpm}$, $129\,\mathrm{rpm}$, $182\,\mathrm{rpm}$, $237\,\mathrm{rpm}$, and $292\,\mathrm{rpm}$) in our experimental setup. The positive and negative thresholds are set to $20\,\%$.

7.3. HS-ERGB Dataset

The HS-ERGB dataset is captured by a dual camera setup [5]. As the CIS and EVS pixel coordinates are not well-aligned due to the configuration of stereo cameras, some of the pixels on the CIS sensor plane lack corresponding EVS data. The pixel-wise reconstruction methods suffer from this issue and generate frames with "holes". To address this issue, we employ a post-processing methodology to yield "hole-free" event data. This method accumulates events in a single frame, followed by a morphological closing operation. The difference between the original accumulated frame and the closed-operated frame serves as the mask for identifying hole pixels. Subsequently, a median filter with a 3×3 kernel is applied to the masked region to fill holes. Note that for machine learning-based methods (Time Lens, FILM, CBMNet, EvUnroll, and REFID), the post-processing method is not applied since the neighborhood relationship is intrinsically considered during the reconstruction of each pixel.

8. Parameter Settings

We use the following settings for sensor-related parameters: $\tau_{0\mathrm{up}} = 3.1\times10^{-5}$, $\tau_{0\mathrm{down}} = 3.2\times10^{-5}$, $\alpha_{\mathrm{up}} = 3.0\times10^{-16}$ and $\alpha_{\mathrm{down}} = 6.5\times10^{-16}$ (for positive and negative events) stem from design. The contrast threshold ranges from $16\,\%$ to $23\,\mathrm{percent}$, and $t_{\mathrm{rp}} = 20\,\mu\mathrm{s}$ comes from hardware settings.

9. Additional Results

The optimization problem is non-linear, non-convex and there is no general theoretical guarantee of convergence. We empirically verified that we converge to reasonable accuracy as can be seen in our benchmark. The numerical optimization is performed using Trust Region Reflective algorithm [2]. We define lower and upper variable boundaries helping stability and use the refinement network to smooth unstable pixels.

Figure 12B illustrates the reconstructed image quality of the simulation dataset. The "basketball" and "slingshot egg" scenes are analyzed in the main paper. Regarding the checkerboard scene, our results exhibit the cleanest, sharpest, and straightest edges, while other methods suffer from varying degrees of ghosting and RS effects. Although EvUnroll demonstrates commendable effectiveness in rectifying rolling shutter distortions within the pattern, residual ghosting artifacts persist notably at the lower section of the checkerboard, which are also evident on the hand in the "basketball" dataset. In the "running man" dataset, our result demonstrates significantly fewer ghosting artifacts compared to EDI and AKF. In this slow-motion scenario, the V_{FE} delay is small, allowing the learning-based method to maximize its advantages. Our method achieves comparable results to the learning-based method.

Figure 13B illustrates the reconstructed image quality of natural scenes within the measured dataset. The proposed method successfully removes ghosting being more impactful to the overall visual experience. In the "mono drone" and "color drone" scenes, our advantage is limited due to slow object motion. The proposed method exhibits fewer ghosting artifacts on the rotating blade compared to EDI and AKF. Moreover, it achieves a notably sharper reconstruction of the upper-left blade pattern in the "mono drone" scene compared to CBMNet's result. In the "table tennis" scene, our method and FILM demonstrate superior performance, while EDI and AKF suffer from pronounced ghosting. Additionally, Time Lens and CBMNet show some distortions in the alphabet characters. Regarding the "badminton" scene, our results present a higher fidelity in reconstructing the shuttlecock compared to FILM, and the racket compared to CBMNet, EvUnroll, and REFID.

Figure 14 presents reconstructed frames for controlled scenes of the measured dataset. In the "$1000\,\mathrm{lx}$, $237\,\mathrm{rpm}$" scene, the results obtained from learning-based methods show either quantization errors or RS effects. In the "$1000\,\mathrm{lx}$, $292\,\mathrm{rpm}$" scene, the reconstructed edges of spokes from learning-based methods begin to generate distortions, and the quality deteriorates as the radial distance increases. Conversely, our method maintains stable performance in these aspects.

Figure 15 displays comparative results for selected scenes in the HS-ERGB dataset. As explained in the main paper, our method does not yield optimal image quality, since it can be influenced by sensor characteristics, data quality (such as the

alignment of CIS and EVS), and object motion. Despite differing sensor characteristics from our model, these results show the excellent generalization capability of our method. Through the comprehensive comparative analysis, it is evident that these SOTA methodologies exhibit imperfections in at least one of the following categories: color noise, blur, or ghosting. In this study, we propose a novel approach that offers a favorable balance between these competing factors.

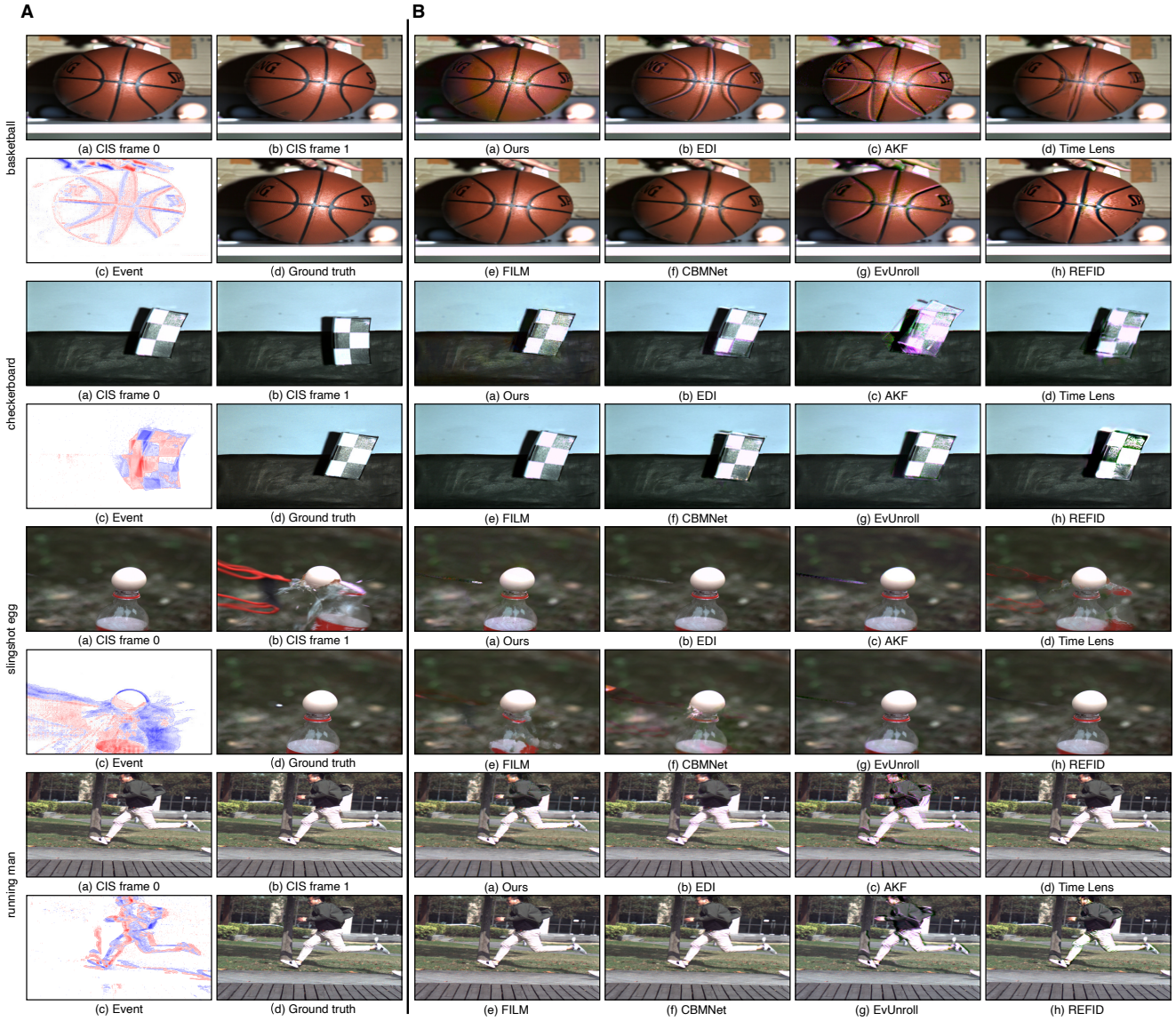

Figure 12. **A**: Visualization of simulation data. "CIS frame 0" and "CIS frame 1" are two consecutive simulated CIS frames that are generated from the simulator; "Event" shows the spatial distribution of generated events whose timestamps are between these two CIS frames (blue: positive events, red: negative events, color saturation: event count number); "Ground truth" shows high frame rate images that are used as the input of the simulator and the evaluation metric calculation. Both CIS frames and events are fed into our joint deblurring, rolling-shutter correction and video frame interpolation method. **B**: Qualitative results comparing the proposed method with EDI, AKF, Time Lens, FILM, CBMNet, EvUnroll, and REFID on the simulation dataset.

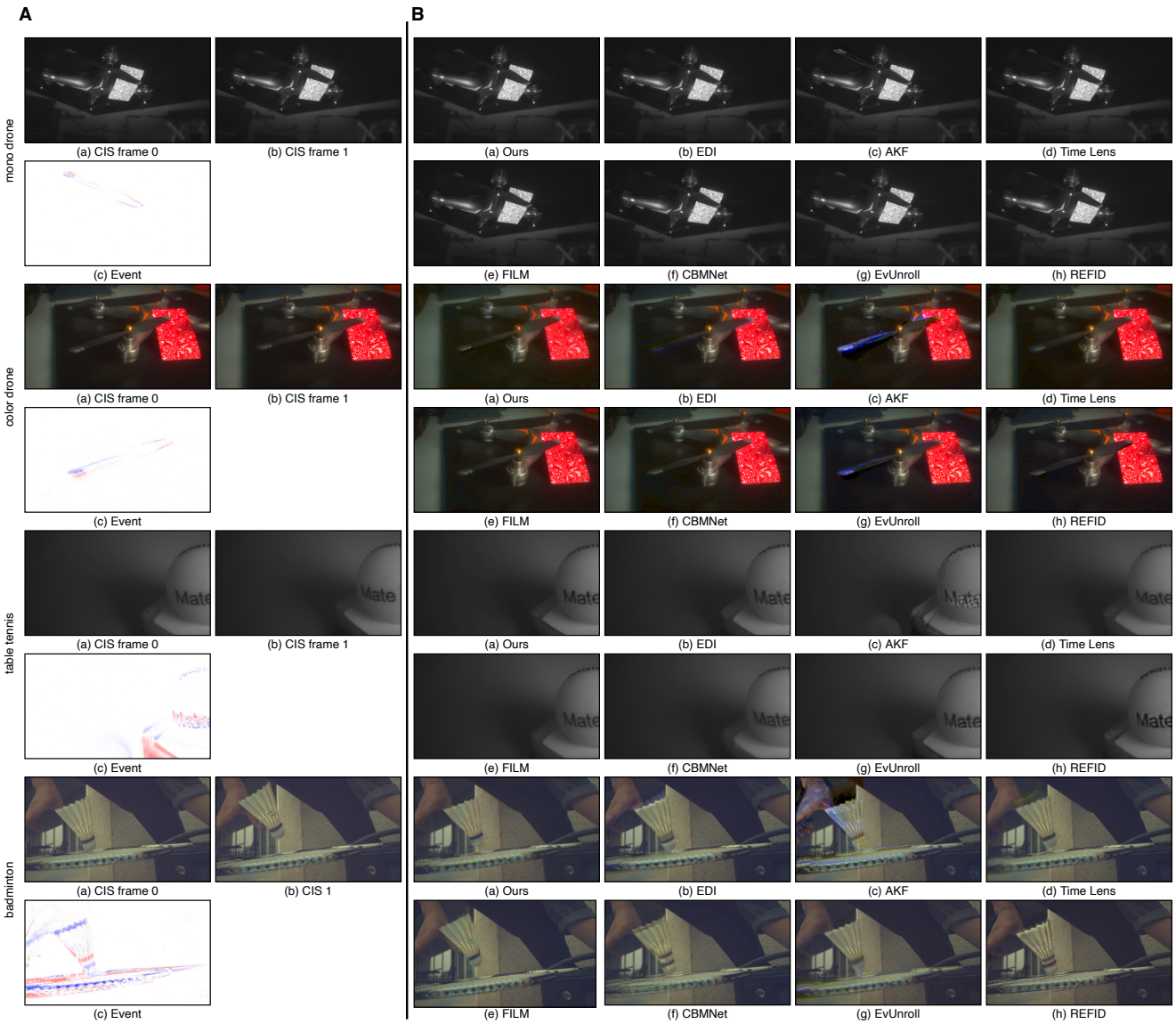

Figure 13. **A**: Visualization of measured data in the natural environment. "CIS frame 0" and "CIS frame 1" are two consecutive CIS frames that are captured by the hybrid sensor; "Event" shows the spatial distribution of collected events whose timestamps are between these two CIS frames (blue: positive events, red: negative events, color saturation: event count number). Both CIS frames and events are fed into our proposed method. GT is not available in the measurement dataset. **B**: Qualitative results comparing the proposed method with EDI, AKF, Time Lens, FILM, CBMNet, EvUnroll, and REFID on natural scenes from the measured datasets.

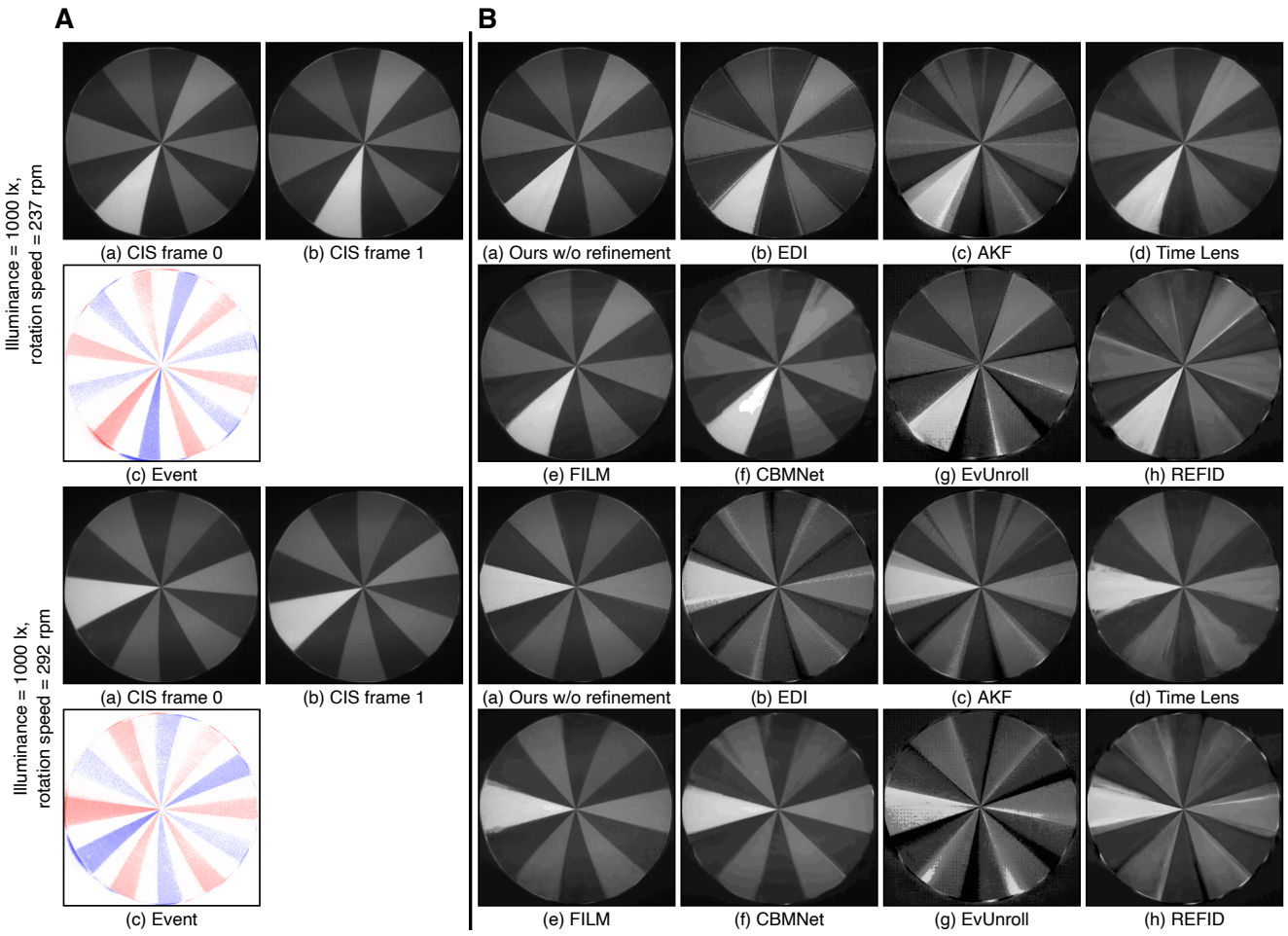

Figure 14. **A**: Visualization of measured data in the natural environment. "CIS frame 0" and "CIS frame 1" are two consecutive CIS frames that are captured by the hybrid sensor; "Event" shows the spatial distribution of collected events whose timestamps are between these two CIS frames (blue: positive events, red: negative events, color saturation: event count number). Both CIS frames and events are fed into our proposed method. GT is not available in the measurement dataset. **B**: Qualitative results comparing the proposed method with EDI, AKF, Time Lens, FILM, CBMNet, EvUnroll, and REFID on controlled scenes from the measured datasets.

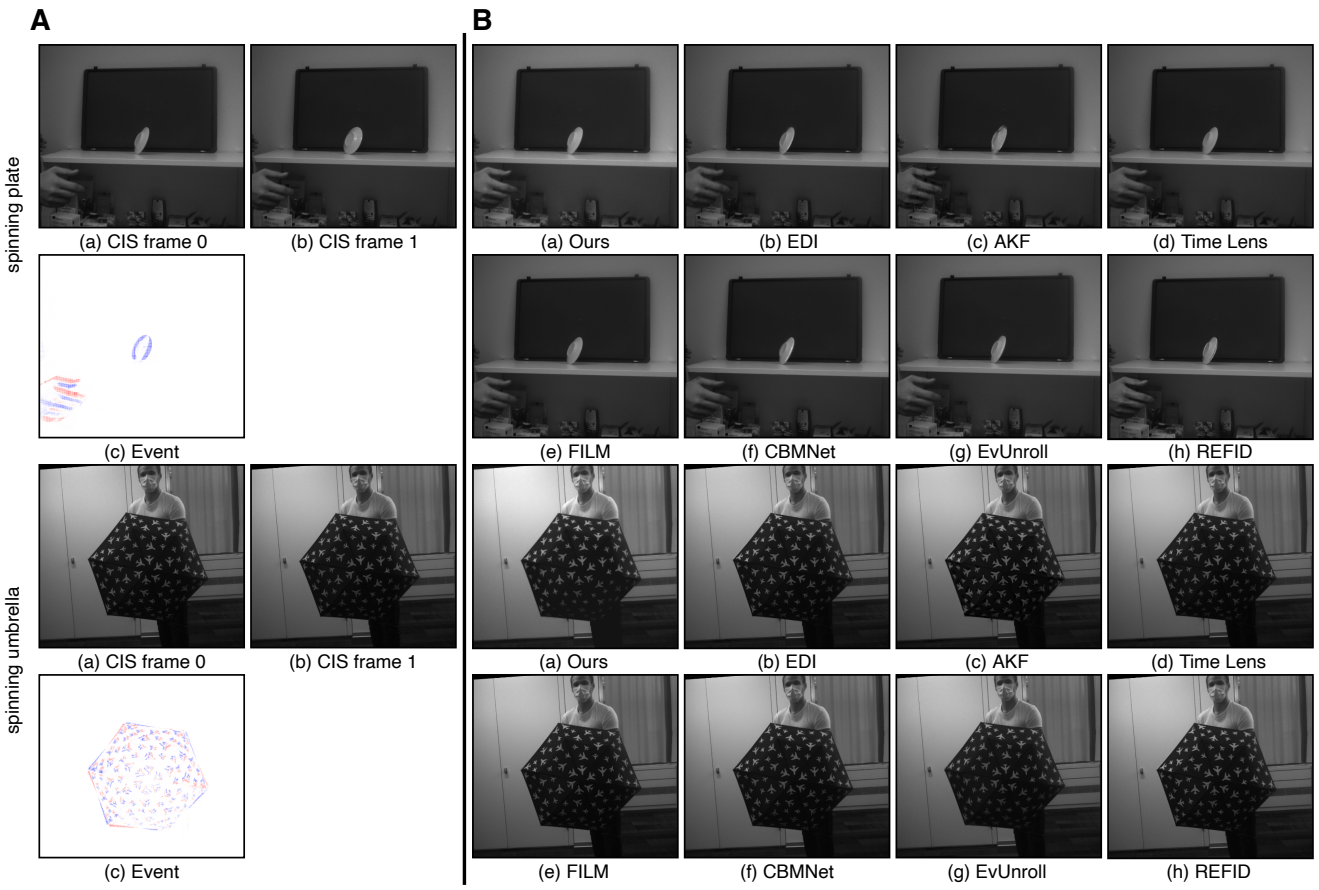

Figure 15. **A**: Visualization of measured data in the natural environment. "CIS frame 0" and "CIS frame 1" are two consecutive CIS frames that are captured by the hybrid sensor; "Event" shows the spatial distribution of collected events whose timestamps are between these two CIS frames (blue: positive events, red: negative events, color saturation: event count number). Both CIS frames and Events are fed into our proposed method. **B**: Qualitative results comparing the proposed method with EDI, AKF, Time Lens, FILM, CBMNet, EvUnroll, and REFID on selected scenes from the HS-ERGB dataset.

10. Additional Ablation Results

Figure 16 demonstrates the ablation results for the "slingshot egg" scene. In this particular scene, the pellet's shape distinctly emerges upon implementing V_{FE} delay compensation. On the other hand, RL and RP compensation yield slight improvements on the pellet's appearance. This observation suggests that V_{FE} plays a more significant role in this instance. Conversely, in the "basketball" scene displayed in the main paper, each module addresses specific aspects of ghosting. The ablation study shows that the efficacy of compensation is scene-dependent, and the proposed compensation methods are indispensable in improving reconstructed image quality.

Figure 17 and Table 4 present an ablation study against NAFNet [1] for deblurring on CIS images. Numerically, the proposed method achieves higher scores on most of the datasets. Visually it is apparent that for the basketball scene our method can reconstruct a more detailed texture on the basketball surface. NAFNet leads to patch artifacts near the image center. The artifacts are also noticeable in the checkerboard scene on black areas in both background and foreground. The study confirms that our method, with the assistance of EVS, exhibits excellent stability in deblurring task compared to NAFNet.

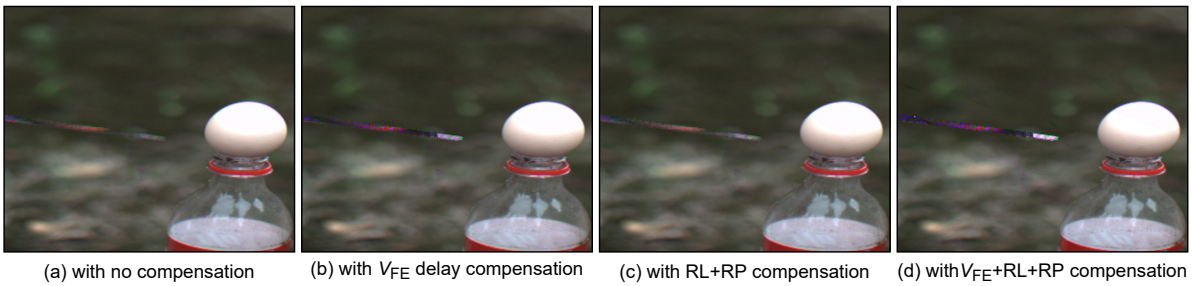

Figure 16. Qualitative results of ablation study on the "slingshot egg" scene. "RL" and "RP" denote "Readout Latency" and "Refractory Period", respectively. Refer to Figure 1(a) for "with V_{FE}+RL+RP compensation and refinement" result.

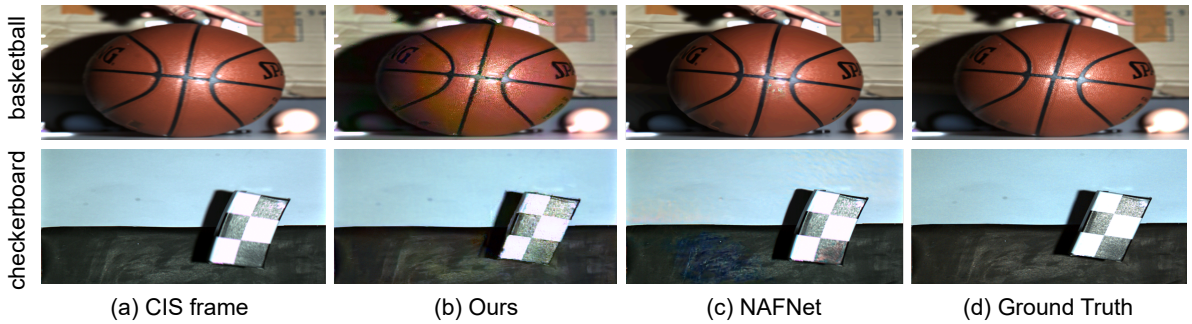

Figure 17. Qualitative comparative results of deblurring with NAFNet.

Table 4. Quantitative comparison of deblurring using original NAFNet [1] on the proposed simulation dataset for ablation study. Each row shows results for a particular scene. The first place is highlighted with **bold underline**.

	Ours		**NAFNet**	
Scene	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
basketball	**27.41**	**0.152**	25.09	0.258
checkerboard	**29.26**	**0.183**	22.27	0.265
slingshot egg	**30.01**	**0.221**	29.61	0.351
running man	**25.98**	0.214	25.51	**0.154**
fan	**24.06**	0.171	19.92	**0.153**
Average	**27.34**	**0.188**	24.48	0.236

References

[1] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 6, 7

[2] MA Branch et al. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, 1999. 2

[3] Menghan Guo, Shoushun Chen, Zhe Gao, Wenlei Yang, Peter Bartkovjak, Qing Qin, Xiaoqin Hu, Dahai Zhou, Qiping Huang, Masayuki Uchiyama, Yoshiharu Kudo, Shimpei Fukuoka, Chengcheng Xu, Hiroaki Ebihara, Xueqing Wang, Peiwen Jiang, Bo Jiang, Bo Mu, Huan Chen, Jason Yang, T. J. Dai, and Andreas Suess. A Three-Wafer-Stacked Hybrid 15-MPixel CIS + 1-MPixel EVS With 4.6-GEvent/s Readout, In-Pixel TDC, and On-Chip ISP and ESP Function. *IEEE Journal of Solid-State Circuits*, pages 1–10, 2023. 2

[4] Xiaozheng Mou, Kaijun Feng, Alex Yi, Steve Wang, Huan Chen, Xiaoqin Hu, Menghan Guo, Shoushun Chen, and Andreas Suess. Accurate event simulation using high-speed video. *Electronic Imaging*, 34(7):242–1–242–6, Jan. 2022. 1

[5] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time Lens: Event-based Video Frame Interpolation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16150–16159, Nashville, TN, USA, June 2021. IEEE. 2