Spatio-Temporal Difference Guided Motion Deblurring with the Complementary Vision Sensor

Supplementary Material

6. Appendix

6.1. Choice of the SD Index

In the main paper (Sec. 3.2, Eq. 2), we mention that our deblurring model takes the $\mathcal{SD}$ frame closest to the exposure midpoint, $\mathcal{SD}_{\lfloor (N-1)/2 \rfloor}$, as structural guidance and outputs the final deblurred image $\mathbf{D}$. Consequently, the recovered structure of $\mathbf{D}$ is aligned with the physically captured structural snapshot $\mathcal{SD}_{\lfloor (N-1)/2 \rfloor}$.

$$\mathbf{D}_k = \mathcal{M}^*\left(\mathbf{B},\, \mathcal{SD}_k,\, \{\mathcal{TD}_i\}_{i=0}^{N-2}\right), \tag{11}$$

where $\mathbf{D}_k$ is structurally aligned with $\mathcal{SD}_k$.

Using the same dataset and training procedure, the network $\mathcal{M}^*$ can thus be optimized to restore the blurry image $\mathbf{B}$ into any temporal slice within the exposure window, i.e., $\mathbf{D}_k$ aligned to $\mathcal{SD}_k$ for any $k \in [0, N-1]$. When applied at inference time, such a model enables the recovery of a short video sequence from a single blurry RGB image and its corresponding spatio–temporal difference signals, depicting the scene motion within the exposure duration.

Full video results are provided in the supplementary material (see `single_frame_to_video.mp4`).

6.2. TD-Sequence Augmentation for Continuous Exposure Time Generalization.

In the main paper (Sec. 3.2, Eq. 2), the input temporal-difference sequence $\{\mathcal{TD}_i\}_{i=0}^{N-2}$ is defined with

$$N = \left\lceil \frac{t_{\text{RGB}}}{\tau_{\text{diff}}} \right\rceil,$$

which guarantees that all motion cues occurring within the RGB exposure duration t_{RGB} are fully covered even when t_{RGB} is not an integer multiple of the difference sampling interval τ_{diff}. However, this also implies that the last element $\mathcal{TD}_{N-2}$ may contain extra motion information beyond the actual exposure window represented by the blurry image.

To enable the network to automatically select valid temporal information from a $\mathcal{TD}$ sequence—and thus generalize to *arbitrary continuous exposure times*—we introduce a temporal-augmentation strategy during training. Specifically, we randomly extend the final TD entry by appending additional TD frames sampled from later timestamps. Formally, we replace:

$$\mathcal{TD}_{N-2} \;\longrightarrow\; \mathcal{TD}^{\star}_{N-2} \quad \text{where} \quad \mathcal{TD}^{\star}_{N-2} = \sum_{j=0}^{m} \mathcal{TD}_{N-2+j},$$

with the augmentation length m randomly sampled from $\{1, 2, 3\}$, and frames added in chronological order. This simulates the situation where the final $\mathcal{TD}$ frame contains extra motion beyond the exposure period while requiring the network to learn to extract only the useful temporal cues.

We compare models trained with and without this TD-augmentation strategy, and evaluate them on the original test set, which provides deblurring ground truth under several discrete exposure durations. As shown below, the augmentation causes negligible performance changes while providing significantly improved generalization to continuous exposure time (see Sec. 6.6).

Augment	5		7		9		11	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×	41.88	0.9912	41.47	0.9905	40.72	0.9887	40.12	0.9874
✓	41.88	0.9913	41.46	0.9905	40.68	0.9887	40.05	0.9873

The negligible difference between the two configurations demonstrates that the augmented model successfully learns to adaptively select valid temporal information, enabling generalization to arbitrary continuous exposure durations.

6.3. DMD-based Dataset Making Details

Existing datasets for novel sensors fall into two categories. Simulation-based methods use software simulators (e.g., ESIM for event cameras) [60] and conveniently reuse existing RGB datasets, but they fail to model real sensor behaviors such as noise, latency, nonlinear response, and refractory effects, limiting their real-world generalization [64]. Real-capture datasets using beam splitters [53–56], dual cameras [51], or motorized rails [61] better capture real sensor responses but often suffer from spatial misalignment, uneven light-intensity distribution, and restricted scene diversity, making large-scale acquisition difficult.

We employ a Digital Micro-Mirror Device (DMD) and its corresponding optical path to project light onto the CVS sensor, enabling the acquisition of realistic sensor responses. We next detail how this mechanism converts an existing high-frame-rate dataset (hereafter referred to as the RGB source, e.g., SportsSloMo [50]) into a CVS deblurring dataset.

We first describe how a single RGB frame is projected onto the CVS: A constant-intensity white light source is reflected by the DMD and projected onto the CVS sensor. By controlling the on/off state of each micro-mirror, we modulate the duration for which each CVS pixel is exposed to the

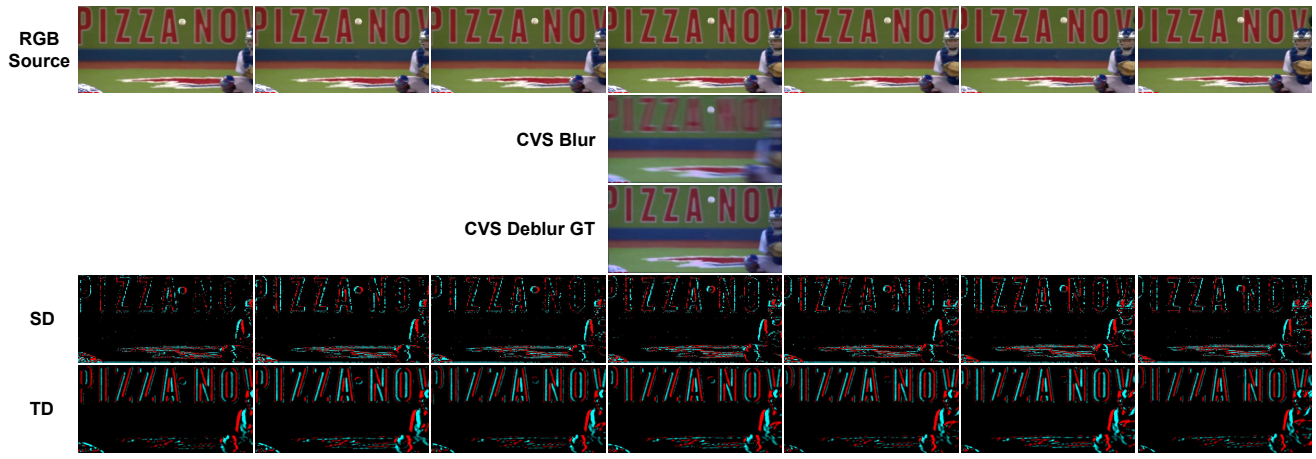


Figure 7. From high-frame-rate SportsSloMo RGB source to the generated SportsSloMo-CVS dataset.

illumination. Assuming the maximum light transmittable duration is T_0, the RGB source pixel at (x, y) has value a, and the maximum pixel value is A, we set the mirror's on-time to $(a/A), T_0$. This produces a physically valid response to the projected image. Because the DMD operates in grayscale, we sequentially project the R, G, and B channels and then combine the CVS imaging results to reconstruct the full RGB image.

To generate temporally varying CVS data—including blurred RGB exposures and real spatio-temporal difference responses—we project a high-frame-rate RGB sequence onto the DMD and synchronize it with the CVS. Specifically, each exposure start signal of the CVS spatio-temporal-difference frame triggers the DMD to display the next RGB source frame. After a duration of τ_{diff}, the CVS issues the next exposure start signal, and the DMD advances to the next frame. Fig. 7 illustrates the projected RGB source frames and the resulting CVS spatio-temporal difference signals.

During this process, the CVS RGB exposure time is adjustable and typically spans multiple consecutive DMD-projected frames. As a result, several sharp source frames overlap within a single exposure, naturally producing realistically blurred RGB images. If we instead repeatedly project a static RGB frame during the exposure, the CVS captures a sharp, blur-free results, which we use as the ground-truth (GT) reference for deblurring.

6.4. Full Comparisons with Other Methods on Real-World Data

In the main paper (Sec. 4.3), we evaluate the real-world generalization ability of different sensor–algorithm combinations. Here, we provide the complete test results, covering both CVS-based pipelines and event-camera-based pipelines.

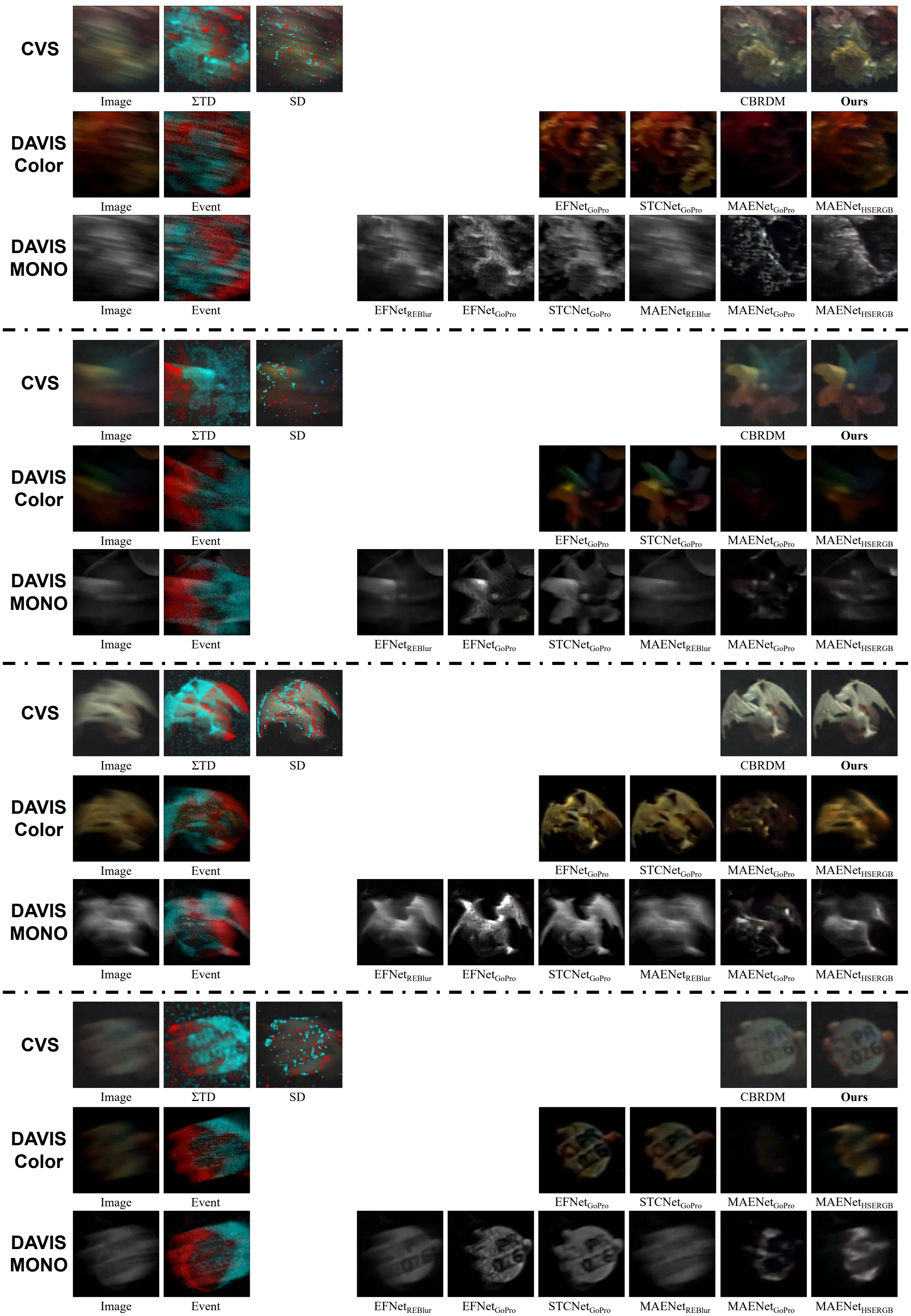
Table 3. All sensor–algorithm–pretraining combinations evaluated on real-world data.

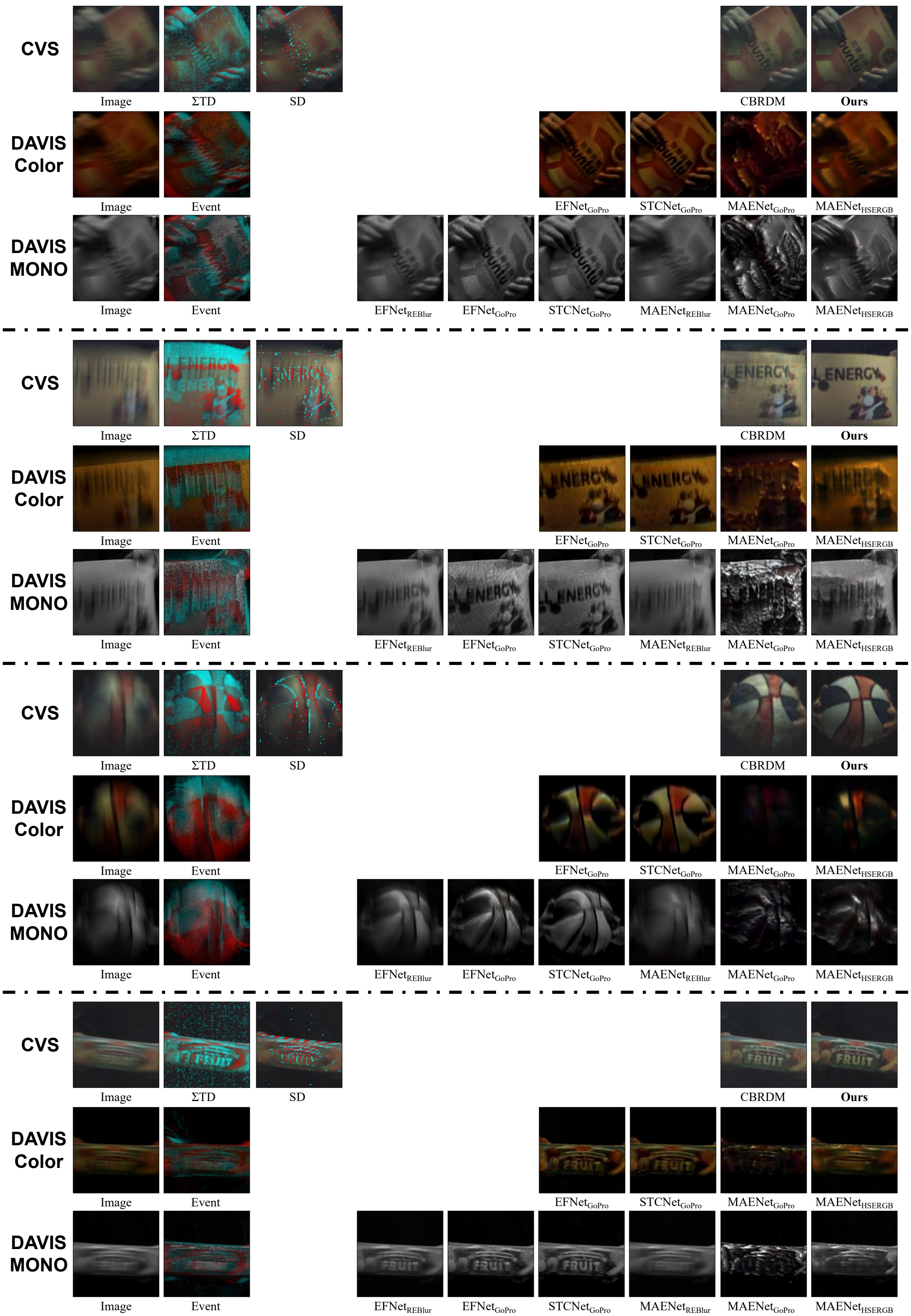
Sensor	Method	Pretrained Dataset
CVS	CBRDM	SportsSloMo-CVS
DAVIS-MONO	EFNet	REBlur
DAVIS-MONO	EFNet	GoPro
DAVIS-Color	EFNet	GoPro
DAVIS-MONO	STCNet	GoPro
DAVIS-Color	STCNet	GoPro
DAVIS-MONO	MAENet	REBlur
DAVIS-MONO	MAENet	GoPro
DAVIS-MONO	MAENet	HSERGB
DAVIS-Color	MAENet	GoPro
DAVIS-Color	MAENet	HSERGB

We include the state-of-the-art CVS reconstruction method *CBRDM* [59], and event-based deblurring models *EFNet* [61], *STCNet* [63], and *MAENet* [62]. For methods offering multiple pretrained variants, we evaluate all applicable weights. Following a unified protocol:

- Models pretrained on grayscale datasets are evaluated only on *DAVIS-MONO*.
- Models pretrained on color datasets are evaluated on both *DAVIS-MONO* and *DAVIS-Color*.

Table 3 lists all sensor–model–pretraining combinations included in our real-world comparison. The main paper reports only representative scenes and the best-performing results for each method; here we provide the full results in Fig. 8 of this supplementary material.


CVS
Image
ΣTD
SD
CBRDM
Ours
DAVIS Color
Image
Event
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
DAVIS MONO
Image
Event
$EFNet_{REBlur}$
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{REBlur}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
CVS
Image
ΣTD
SD
CBRDM
Ours
DAVIS Color
Image
Event
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
DAVIS MONO
Image
Event
$EFNet_{REBlur}$
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{REBlur}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
CVS
Image
ΣTD
SD
CBRDM
Ours
DAVIS Color
Image
Event
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
DAVIS MONO
Image
Event
$EFNet_{REBlur}$
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{REBlur}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
CVS
Image
ΣTD
SD
CBRDM
Ours
DAVIS Color
Image
Event
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
DAVIS MONO
Image
Event
$EFNet_{REBlur}$
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{REBlur}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$


CVS
Image
ΣTD
SD
CBRDM
Ours
DAVIS Color
Image
Event
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
DAVIS MONO
Image
Event
$EFNet_{REBlur}$
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{REBlur}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
CVS
Image
ΣTD
SD
CBRDM
Ours
DAVIS Color
Image
Event
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
DAVIS MONO
Image
Event
$EFNet_{REBlur}$
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{REBlur}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
CVS
Image
ΣTD
SD
CBRDM
Ours
DAVIS Color
Image
Event
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
DAVIS MONO
Image
Event
$EFNet_{REBlur}$
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{REBlur}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
CVS
Image
ΣTD
SD
CBRDM
Ours
DAVIS Color
Image
Event
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$
DAVIS MONO
Image
Event
$EFNet_{REBlur}$
$EFNet_{GoPro}$
$STCNet_{GoPro}$
$MAENet_{REBlur}$
$MAENet_{GoPro}$
$MAENet_{HSERGB}$

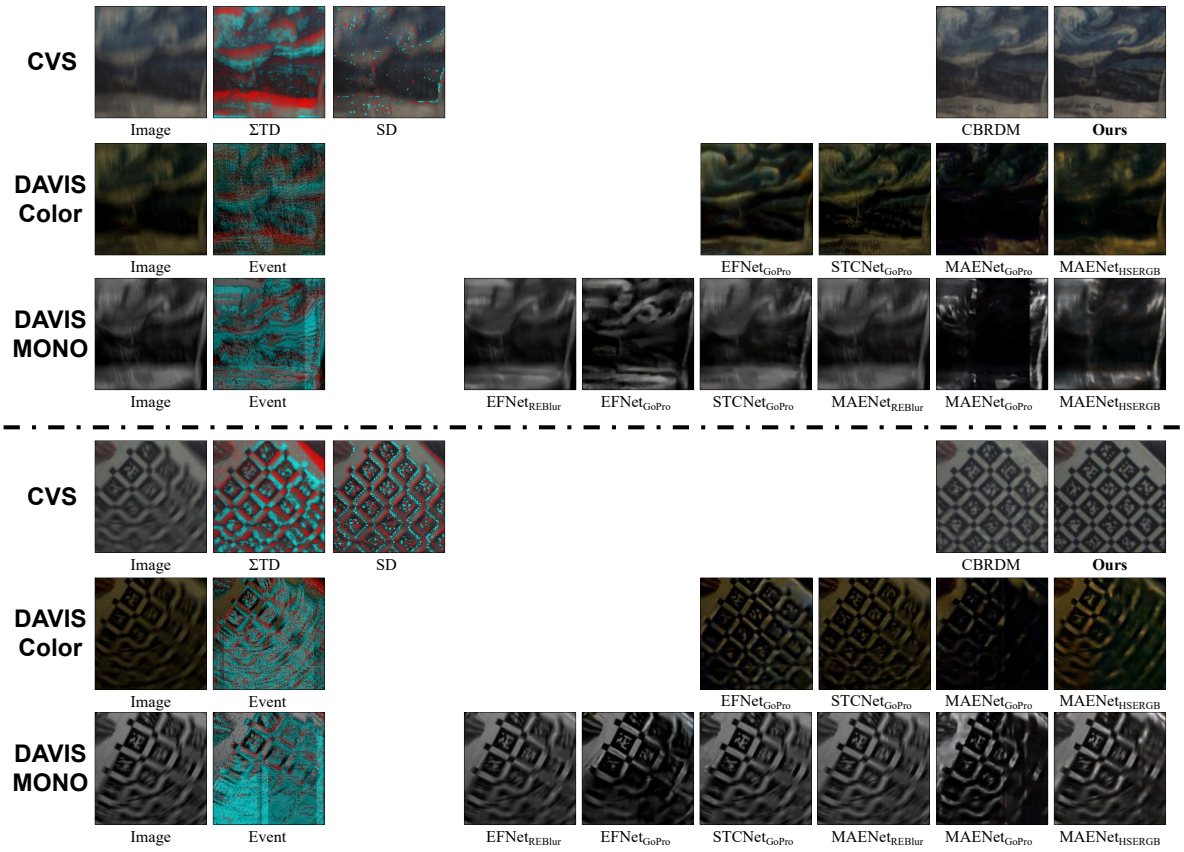


Figure 8. Full comparison results on real-captured data with CVS-based and event-based methods.

6.5. Performance Boundary Analysis Details

To investigate the performance limits of CVS-based motion deblurring in real-world settings, we construct a standardized rotating-disk benchmark that enables controlled, repeatable, and quantitatively measurable evaluation. The setup allows precise adjustment of two key physical factors governing motion blur: (1) the disk rotation speed and (2) the RGB exposure duration.

Data Acquisition. A motor-controlled rotating disk with adjustable angular velocity is illuminated by a tunable lighting system (400–900 lux), yielding CVS RGB exposure times ranging from 6,600 µs to 14,520 µs. The CVS camera is positioned orthogonally to the disk surface, and its pose remains fixed throughout all recordings. For each exposure duration, the disk speed is uniformly varied from 200 rpm to 1000 rpm. In addition to the blurred measurements, we also capture a static, blur-free reference image under the same illumination conditions, which serves as the normalization baseline for subsequent analysis.

Deblurring Evaluation. Because ground-truth sharp frames are unavailable in real-captured experiments, we follow the edge-sharpness-based evaluation procedure of [52]. For each deblurred image produced by ours STGDNet, we uniformly sample points along the disk perimeter and collect their grayscale values to obtain a 1D angular intensity sequence for quantifying residual blur.

1. The disk center is manually annotated. A sampling circle of radius $R/2$ (with R the disk radius) is drawn.
2. Let $\mathbf{p}_0$ denote the intersection between the sampling circle and the black–white boundary. We shift $\mathbf{p}_0$ counterclockwise by $\pi/12$ to align the sampling origin with the center of a black sector.
3. Start from $\mathbf{p}_0$, we uniformly sample $K = 3600$ points along the circle. Let $\theta \in [0, 2\pi)$ be the angular coordinate, yielding the grayscale sequence:
$$g(\theta_k), \quad k = 1, \ldots, K. \tag{12}$$
4. Each transition between two adjacent color sectors spans $\Delta\theta = \pi/6$. For every such interval, we fit a sigmoid function to model the edge transition:
$$S(\theta) = \frac{\Delta}{1 + \exp[-(a\theta + b)]} + g_{\min}, \tag{13}$$

where $\Delta = g_{\max} - g_{\min}$. Parameters (a, b) are estimated using Levenberg–Marquardt optimization [57, 58] (10k iterations).

Blurred Edge Width (BEW). For each fitted sigmoid function, we compute its blurred edge width (BEW) using the 10%–90% intensity transition in the angular domain:

$$S(\theta_{10}) = g_{\min} + 0.1\Delta, \qquad S(\theta_{90}) = g_{\min} + 0.9\Delta, \quad (14)$$

$$\text{BEW} = \theta_{90} - \theta_{10}. \quad (15)$$

Mean Relative BEW. To normalize the blur level across the transitions between different color sectors, we divide each BEW computed from the deblurred image by the corresponding BEW measured from the static reference image:

$$\text{rBEW}_i = \frac{\text{BEW}_i^{(\text{deblur})}}{\text{BEW}_i^{(\text{static})}}, \quad (16)$$

where i indexes the i-th color-transition edge on the disk, and rBEW denotes the relative BEW. The final metric is the mean relative BEW (Mean-rBEW) over all color-transition edges:

$$\text{Mean-rBEW} = \frac{1}{N}\sum_{i=1}^{N} \text{rBEW}_i. \quad (17)$$

Lower Mean-rBEW indicates sharper transitions and therefore higher-quality deblurring relative to the physically sharp reference.

Performance Boundary Visualization. For each exposure duration and rotation speed, we compute the Mean-rBEW and plot the corresponding curves in Fig. 9. Analyzing these curves, we observe that under the same illumination and exposure time, a higher rotation speed leads to a larger Mean-rBEW after deblurring, indicating that stronger initial blur makes the deblurring task more challenging. On the other hand, the Mean-rBEW–rotation-speed curves across different exposure durations show no pronounced differences, suggesting that our method is highly robust to varying exposure times and illumination conditions.

We establish this evaluation protocol and will release the real-captured quantitative dataset as a CVS deblurring benchmark, with the aim of providing an effective and practical metric for real-world deblurring assessment.

6.6. Full Generalization Results

In Fig. 10, we present over 100 blur–deblur pairs covering continuous exposure times, diverse indoor and outdoor scenes, both camera and object motion, and a wide range of linear and nonlinear blur patterns. The results demonstrate our method's stable color fidelity, accurate structural restoration, and strong real-world generalization. Each example corresponds to a short multi-frame video clip, with the full demonstrations provided in `100_real_demos.mp4`.

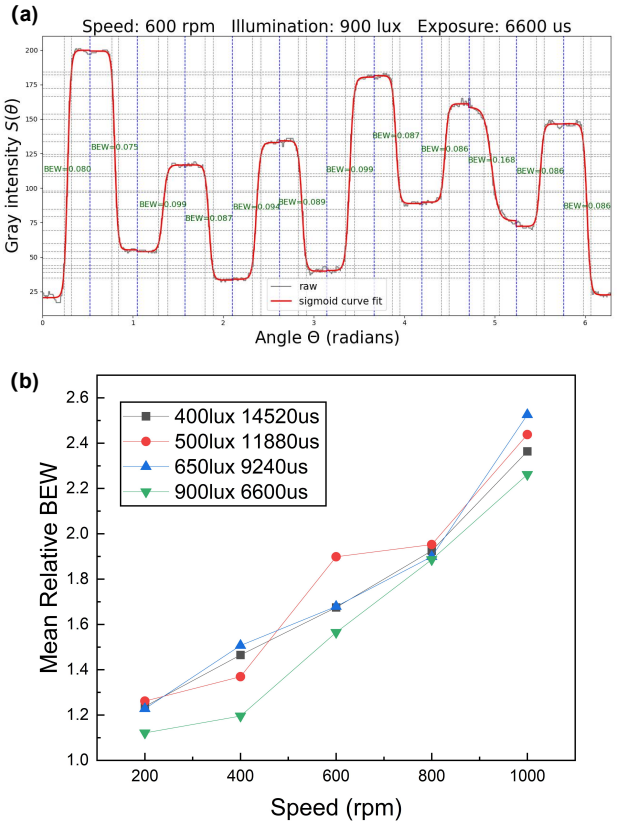


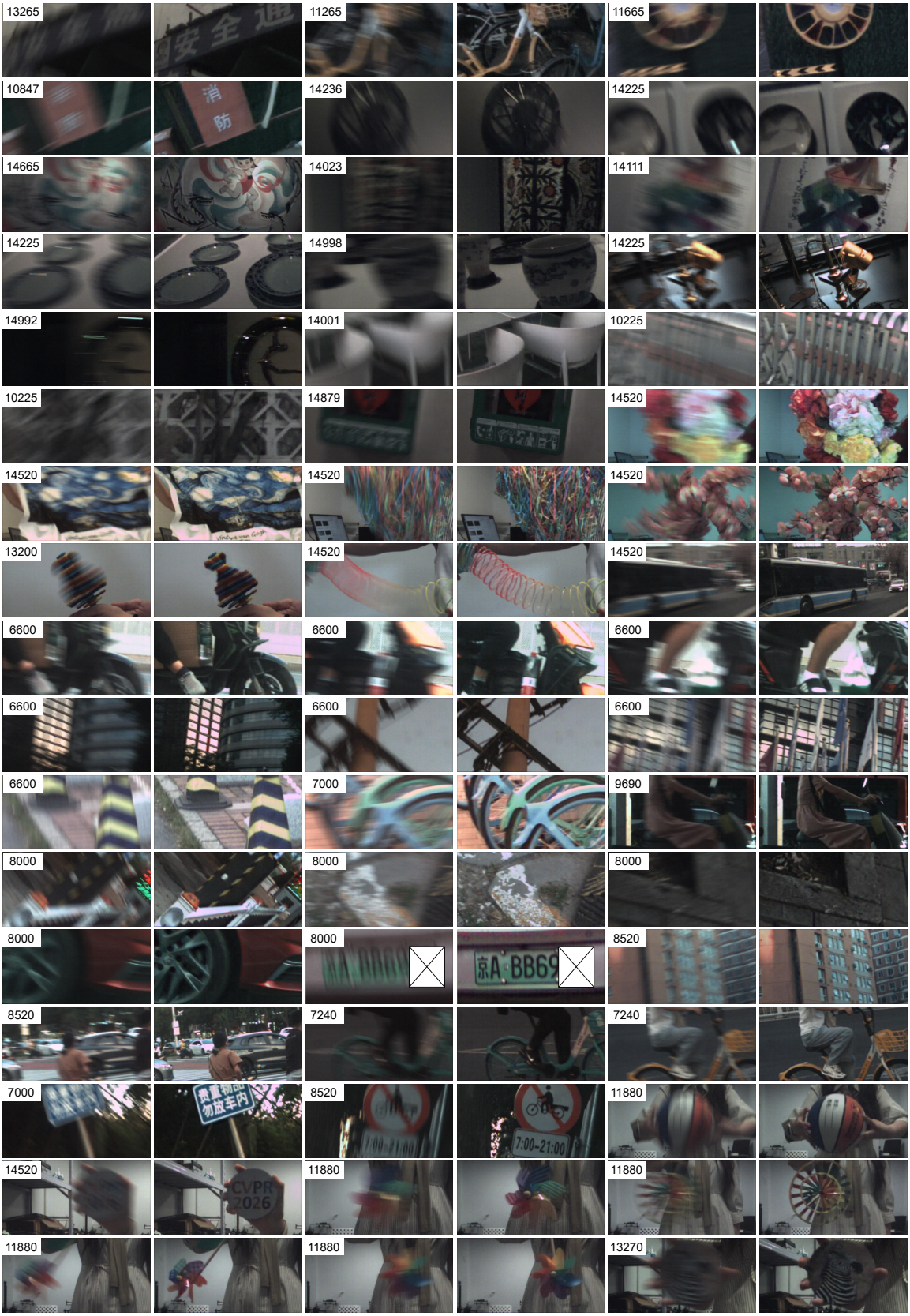
Figure 9. Performance boundary visualization using the rotating-disk benchmark. (a) 1D angular intensity sequence sampled along the disk, together with the corresponding sigmoid fitting results, shown for one example configuration (600 rpm, 900 lux, 6,600 µs exposure). (b) Mean Relative BEW versus rotation speed under different illumination levels.

References

[50] Jiaben Chen and Huaizu Jiang. Sportsslomo: A new benchmark and baselines for human-centric video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6475–6486, 2024.

[51] Hoonhee Cho, Yuhwan Jeong, Taewoo Kim, and Kuk-Jin Yoon. Non-coaxial event-guided motion deblurring with spatial alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12492–12503, 2023.

[52] Hai Dinh, Qinyi Wang, Fangwen Tu, Brett Frymire, and Bo


13265
11265
11665
10847
消
防
14236
14225
14665
14023
14111
14225
14998
14225
14992
14001
10225
10225
14879
14520
14520
14520
14520
13200
14520
14520
6600
6600
6600
6600
6600
6600
6600
7000
9690
8000
8000
8000
8000
8000
8520
8520
7240
7240
7000
8520
11880
14520
11880
11880
11880
11880
13270

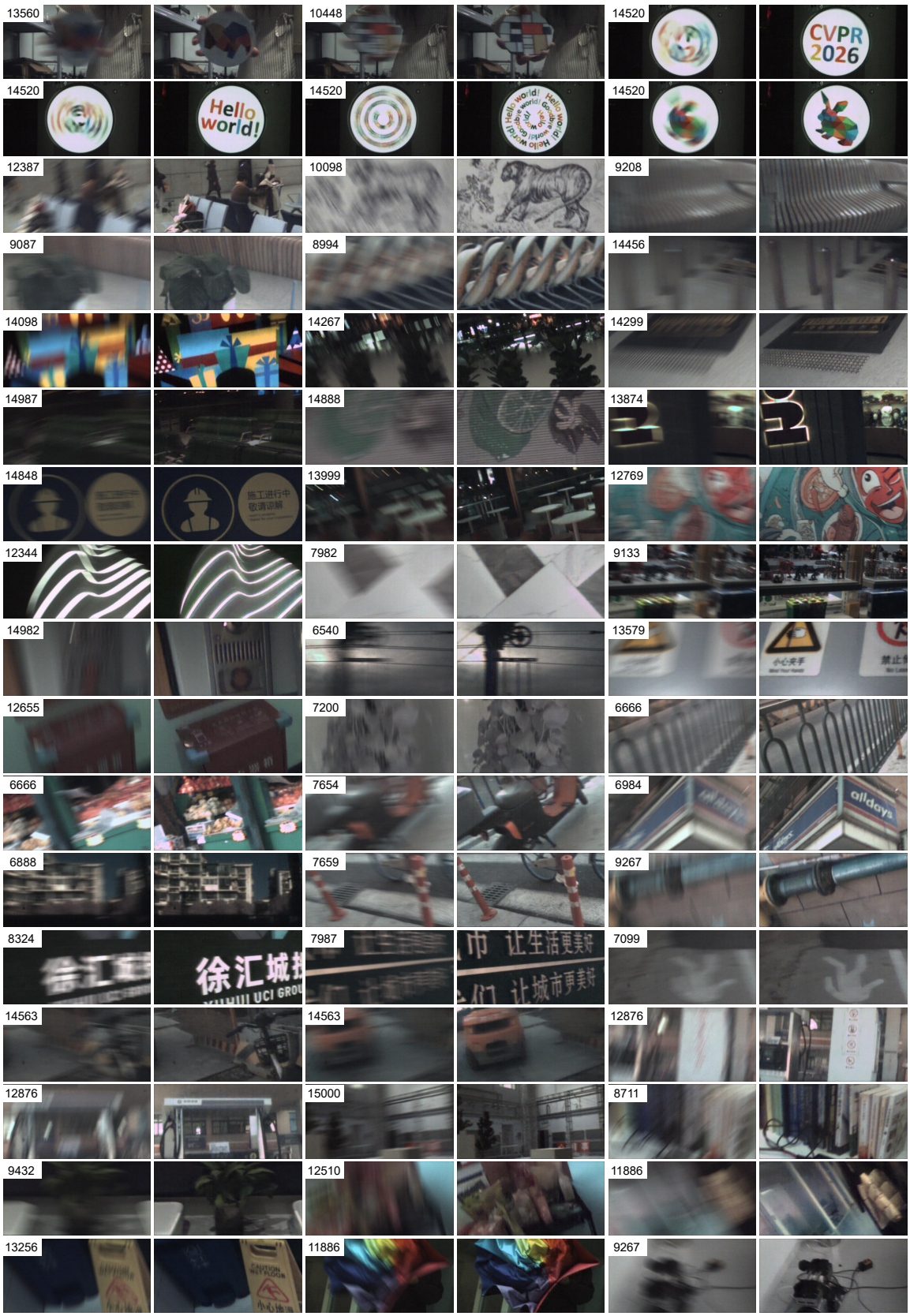


Figure 10. Our method's deblurring results across arbitrary exposure times on 100+ real-world CVS captured scenes.

Mu. Evaluation of motion blur image quality in video frame interpolation. *Electronic Imaging*, 35:262–1, 2023.

[53] Peiqi Duan, Boyu Li, Yixin Yang, Hanyue Lou, Minggui Teng, Xinyu Zhou, Yi Ma, and Boxin Shi. Eventaid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[54] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Cmta: Cross-modal temporal alignment for event-guided video deblurring. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.

[55] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24966–24976, 2024.

[56] Taewoo Kim, Jaeseok Jeong, Hoonhee Cho, Yuhwan Jeong, and Kuk-Jin Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *European Conference on Computer Vision*, pages 433–451. Springer, 2024.

[57] Kenneth Levenberg. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

[58] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[59] Yapeng Meng, Yihan Lin, Taoyi Wang, Yuguo Chen, Lijian Wang, and Rong Zhao. Diffusion-based extreme high-speed scenes reconstruction with the complementary vision sensor. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5701–5710, 2025.

[60] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pages 969–982. PMLR, 2018.

[61] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European conference on computer vision*, pages 412–428. Springer, 2022.

[62] Zhijing Sun, Xueyang Fu, Longzhuo Huang, Aiping Liu, and Zheng-Jun Zha. Motion aware event representation-driven image deblurring. In *European Conference on Computer Vision*, pages 418–435. Springer, 2024.

[63] Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, and Guangming Shi. Motion deblurring via spatial-temporal collaboration of frames and events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6531–6539, 2024.

[64] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022.