Supplemental Material for *Motion Blur Decomposition with Cross-shutter Guidance*

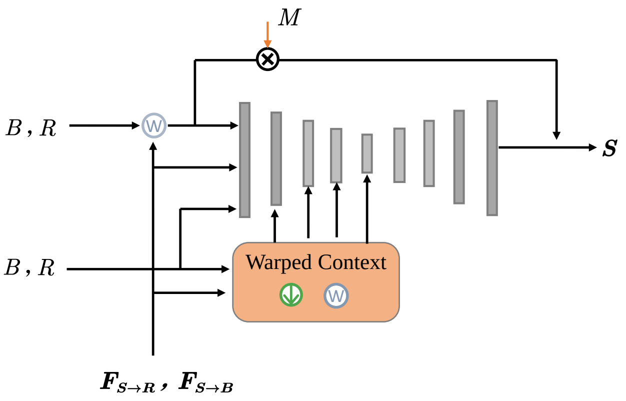
Xiang Ji   Haiyang Jiang   Yinqiang Zheng[†]
The University of Tokyo, Japan
{jixiang,jiang-haiyang777}@g.ecc.u-tokyo.ac.jp,
yqzheng@ai.u-tokyo.ac.jp

In this supplemental material, we present implementation details (Section A) and additional results (Section B) as complements of the main content.

A. Implementation Details

A.1. Architecture Details

The main components of our model have been presented in Section 3. Here we further provide more details about our $GenNet$. Following the design of encoder-decoder with coarse-to-fine scale [3, 16], we implement the $GenNet$ as shown in Figure A.1. $GenNet$ takes as input two original blur and RS views along with motion fields predicted by motion interpretation model. To better refine the warped frames, we also extract the context of blur and RS views with different scale and concatenate them into corresponding encoders. Two groups of warped frames from blur and RS views will be merged through the estimated mask M and then connected to the final outputs.



Extended Figure A.1. **Architecture of the blur decomposition.** To better refine the warped frames, we also exploit the context of blur and RS views to generate temporally and spatially consistent video.

Extended Table A.1. **Specifications of our triaxial imaging system**. The deadtime between two adjacent high speed frames is extremely short and thus can be ignored.

Device	RS camera	GS camera	HS camera
Resolution	800×800	800×800	800×800
Frame rate	20 fps	20 fps	500 fps
Exp. per Row	2 ms	18 ms	2 ms
Delay. per Row	20 μs	0 μs	0 μs
Exp. per Frame	18 ms	18 ms	2 ms
Deadtime	32 ms	32 ms	0 ms

A.2. Construction of Imaging system

Drawing inspirations from recent studies that construct co-axis optical settings, such as [10] and [14], we develop our triaxial imaging system to capture a realistic dataset of strictly aligned RS, blur and high-speed videos. Initially, we fix the RS camera and manually adjust the orientation and position of the other two cameras by evaluating the residual images of a checker pattern. For capturing the same visual content in blur and RS view, the exposure time of a GS camera is elongated to that of RS counterpart. Nonetheless, achieving precise alignment for all three cameras in all directions proved to be exceedingly difficult due to complexity of the system. We then fixed all system components and calibrated the alignment of three cameras by utilizing two homographies, under the consideration of close-loop constraint. Given that two beamsplitters reduced incoming light to a quarter, we collect our dataset on a sunny day to ensure that all images were sufficiently bright. Moreover, we cooled the high-speed camera to ensure the quality of groundtruth frames.The specifications for 3 cameras are detailed in Table A.1.

A.3. Experimental Data and Training Loss

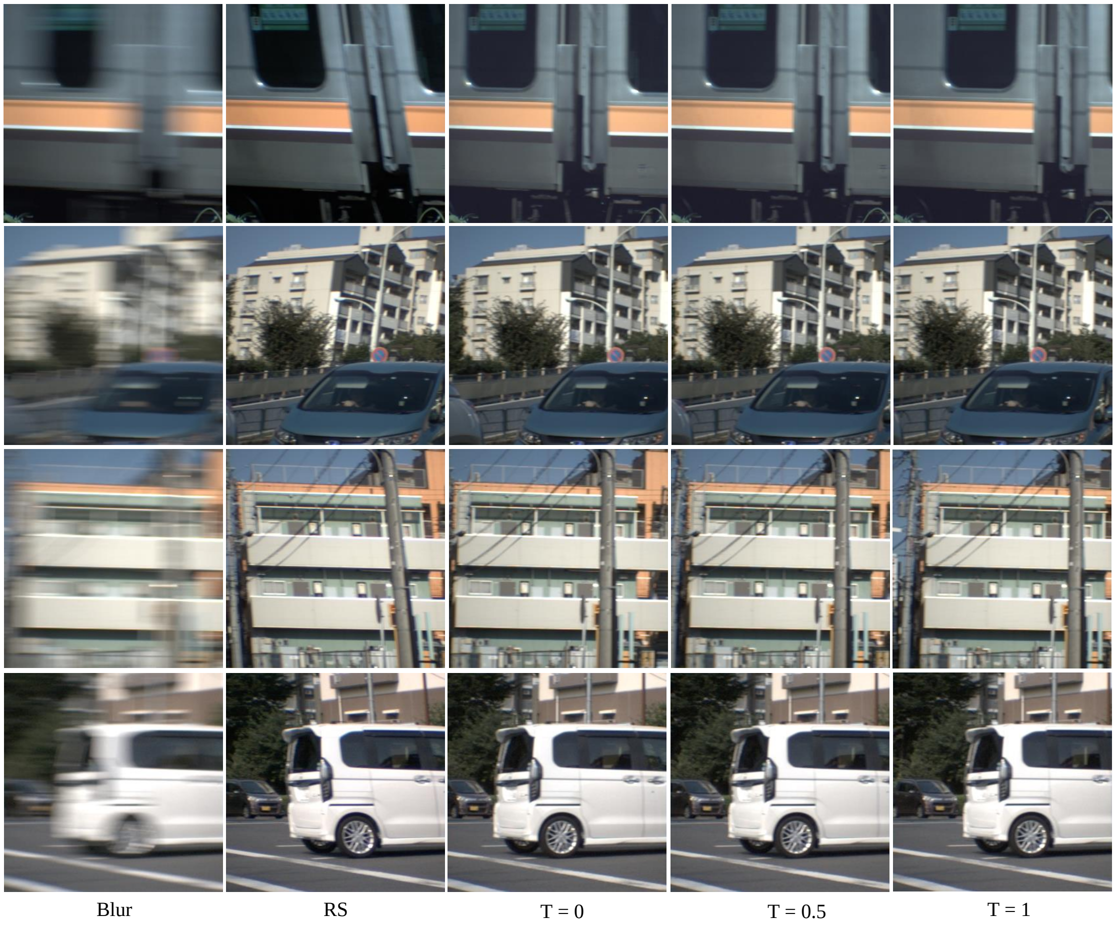
A.3.1 Real Dataset

The detailed collection process of our real data is presented in Section 3.2. Assisted by the imaging system, we established our dataset RealBR including 54 distinct street scenes and each scene has 56 × 3 degraded frames and 1400 sharp HS frames. The presented data samples of RealBR are in

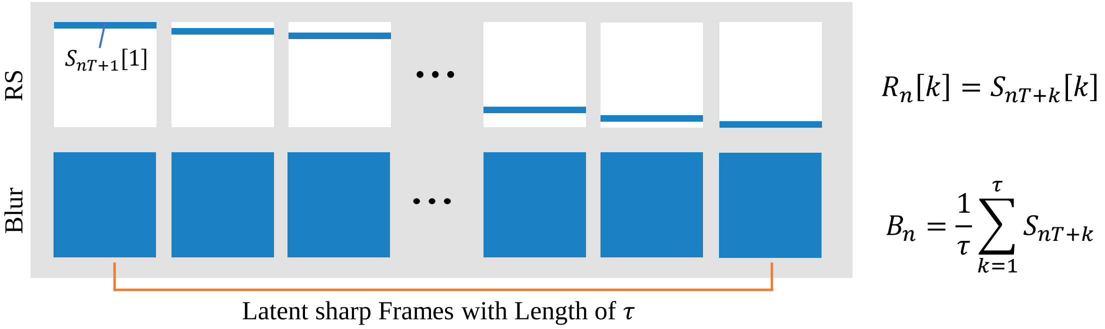
This research was supported in part by JSPS KAKENHI Grant Numbers 22H00529, 20H05951, JST-Mirai Program JPMJMI23G1.

† Corresponding author

https://github.com/jixiang2016/dualBR



Extended Figure A.2. **Data samples from our RealBR**. We present three groundtruth frames which are temporally located at $T = 0, 0.5, 1$.


RS
Blur
$S_{nT+1}[1]$
$R_n[k] = S_{nT+k}[k]$
$B_n = \frac{1}{\tau}\sum_{k=1}^{\tau} S_{nT+k}$
Latent sharp Frames with Length of τ

Extended Figure A.3. **Synthetic method**. The notation $B[k]$ denotes extracting the k-th row from frame B. n is the index of blur or RS view. T is the number of latent frames that correspond to exposure and deadtime.

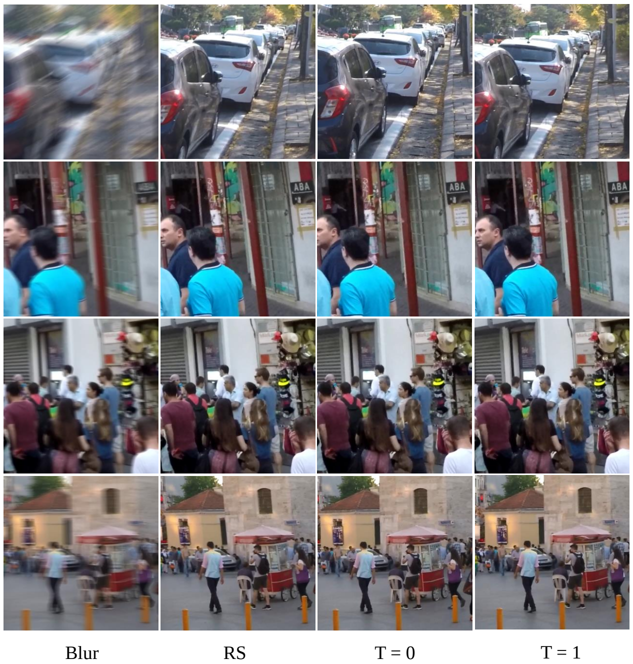
Figure A.2. After necessary preprocessing, we reorganized entire dataset and split it into 40, 4, and 10 scenes for training, validation and test. Notably, the characteristic differences of two views are quite obvious. The RS has local details with tilt effects that encode motion direction of latent frames while blur view contains adequate global context and records relatively precise initial state of objects.

A.3.2 Synthetic Dataset

In order to supplement our findings on the guidance of cross-shutter view to blur decomposition, we additionally construct a synthetic dataset following the protocol in [1, 8, 11]. The synthesizing process is grounded on GOPRO data [7] consisting of 33 videos with resolution 1280 × 720. Each video consists of 1200 consecutive frames captured at 240fps. To achieve more realistic effects, the GOPRO dataset was initially interpolated at a factor of ×64 using an off-the-shelf video interpolation algorithm [3].

Figure A.3 illustrates our synthesizing method. Depicted as in [6, 11], RS videos are generated by sequentially copying a row from corresponding high-speed frames within exposure time τ, and blur observations are synthesized through averaging them. We strictly follow the constraints in Figure A.3 to ensure RS and blur views are aligned in frame level and capture identical content of the scene. Specially, We operate central crop of 512 to each frames, and set $T = \tau = 512$. Figure A.4 exhibits examples of synthesized data.

For comparing with more competitive settings: IFED [16] with *dual reversed RS views*, EvUnroll [18] and EBFI [12] assisted by *event camera*, PMB [9] using *short and long exposure*. We synthesized events from the high frame-rate video using an event simulator [2]. The shortly exposed inputs of PMB are generated by adding Poisson noise to clear latent frames. Moreover, following [16], the synthesizing process of reversed RS view is strictly aligned to original RS and blur views.



Extended Figure A.4. **Data samples from our synthetic data**. We present three groundtruth frames which are temporally located at $T = 0, 1$.

A.3.3 Training Loss

We train our network using the Charbonnier loss [13] to reconstruct clear latent image sequence:

$$\mathcal{L} = \sum_{t=0}^{N-1} \sqrt{\|S^t - G^t\|^2 + \epsilon^2}, \tag{1}$$

where S^t, and G^t denote predicted clear frame and corresponding ground truth at time instance t, respectively. N is the length of reconstructed latent sequence and ϵ is a constant which we empirically set to 10^{-3} for all the experiments.

A.3.4 Implementation Details

As explained in Sec.3.2, each blur-RS pair corresponds to 9 high-speed sharp frames. So, for training and validation, each sample comprises a paired input (B, R) and groundtruth video clip S' with a length of 9. Our model is trained by using Adam optimizer [5] with epoch of 800. The initial learning rate is set to 10^{-4} and decreases to 10^{-6} through a cosine annealing scheduler. To augment the training data, we first crop the samples into 512 and then conduct random horizontal flipping and channel reverse. Experiments are performed on two GPUs of NVIDIA Tesla V100 with batch size of 8. We evaluate the performance of models using standard metrics (PSNR, SSIM and LPIPS). Higher PSNR/SSIM or lower LPIPS suggests better performance. Besides conducting comparisons on our collected real-world dataset RealBR, we also train all models on synthesized dataset based on GOPRO data [7].

B. Additional Results

B.1. Experimental Results on Synthetic Data

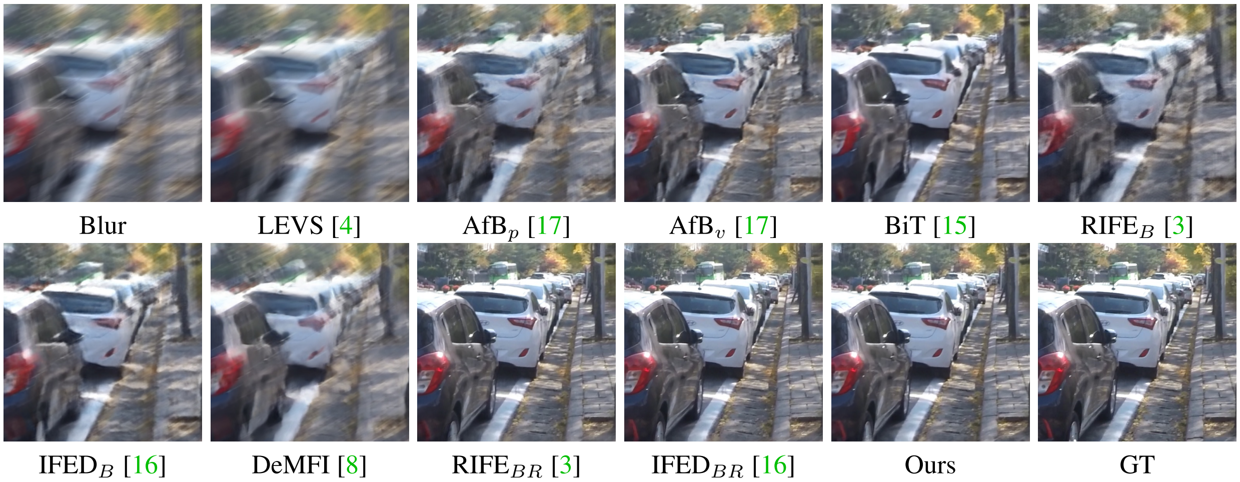
As supplemental demonstration of conclusion drawn on real-data RealBR, we also conduct experiments on synthetic dataset based on GOPRO as shown in the main script. Here we present extra qualitative results in Figure A.5.

Note that, the authors do not provide the training code of LEVS. So, we reproduce training process strictly following the details described in the paper. Due to unknown of some hyper parameters, the performance on our synthetic data is to select the optimal results from: (a) directly using the pre-trained model to infer on our test set; (b) fine-tuning the pre-trained model on our training set then inferring on test set; (c) training from scratch on our data and then inferring on test set.

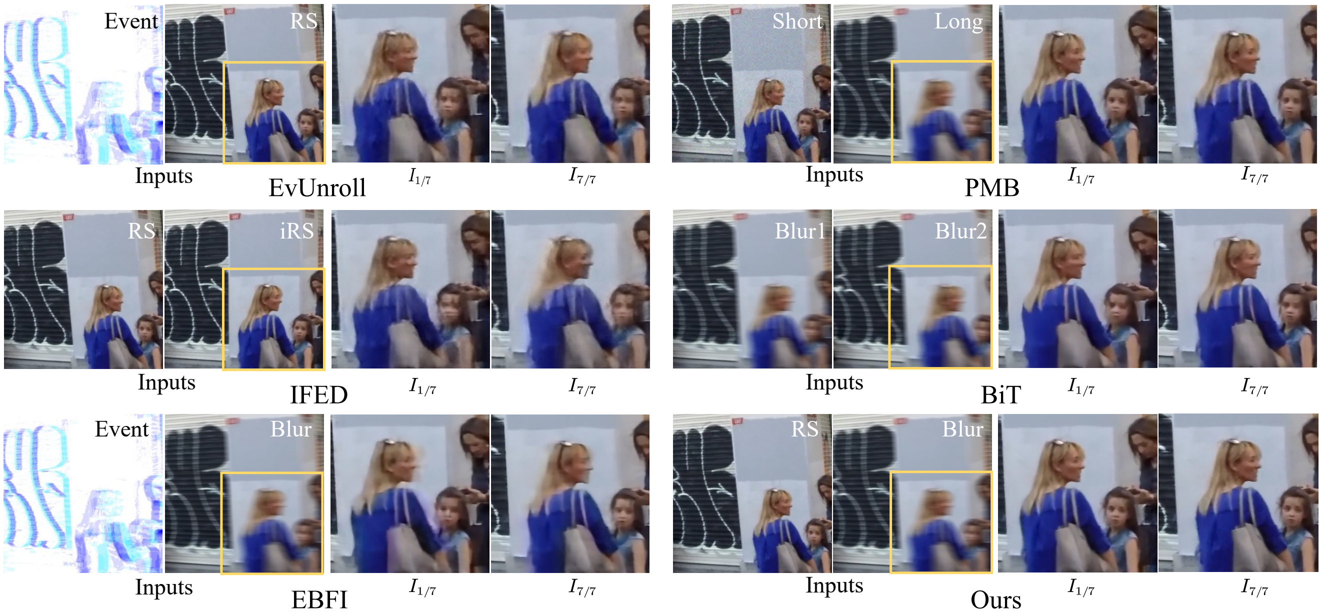
Besides, more visual comparisons with different competitive settings on synthetic data are also provided in Figure A.6 and Figure A.7.

B.2. Video reconstruction results

To substantiate the ability of our model in motion direction disambiguation and local details recovery, we apply all mod-



Extended Figure A.5. **Qualitative comparison on synthetic data.** Our model obviously outperforms the approaches approximating latent motion fields relying on adjacent blurry inputs.



Extended Figure A.6. **Visual results of comparisons with competitive settings** on synthetic data. I_t denotes the interpolated frame temporally located at time instant t.

els to generating 9 latent frames, whose visual results are shown in Figure B.8. In terms of motion direction, LEVS fails to animate the scene, rendering little to none movements in the recovered sight, while ghosting effects, reminiscent of incorrect double exposures, are visible in AfB$_v$, IFED$_B$, BiT, and RIFE$_B$ results, serving as evidences of false motion estimations. As for local details, DeMFI shows an overly smoothed vision, diminishing high frequency characteristics. With the aid of cross-shutter guidance, IFED$_{BR}$, RIFE$_{BR}$, and ours faithfally restore authentic motions, with our results possessing the sharpest details among the three. More experimental results could be found in Figure B.9, Figure B.10 and video demos.

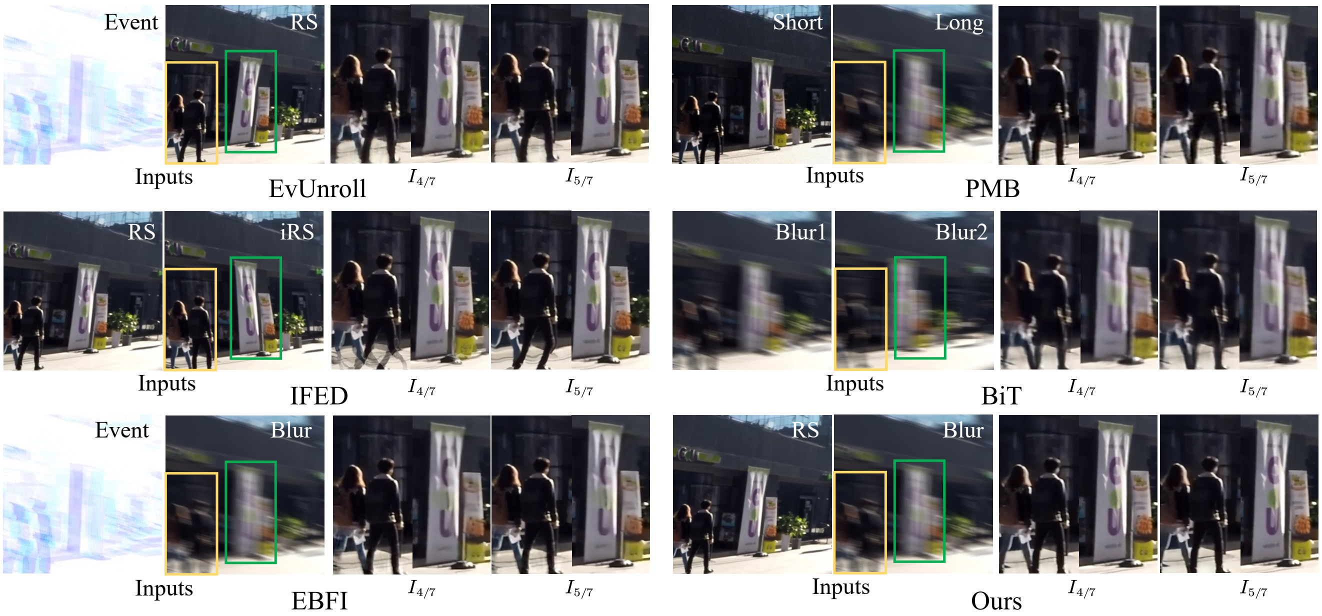
B.3. Additional Qualitative Results

We present additional qualitative comparisons on RealBR dataset in Figure B.11 – Figure B.16.

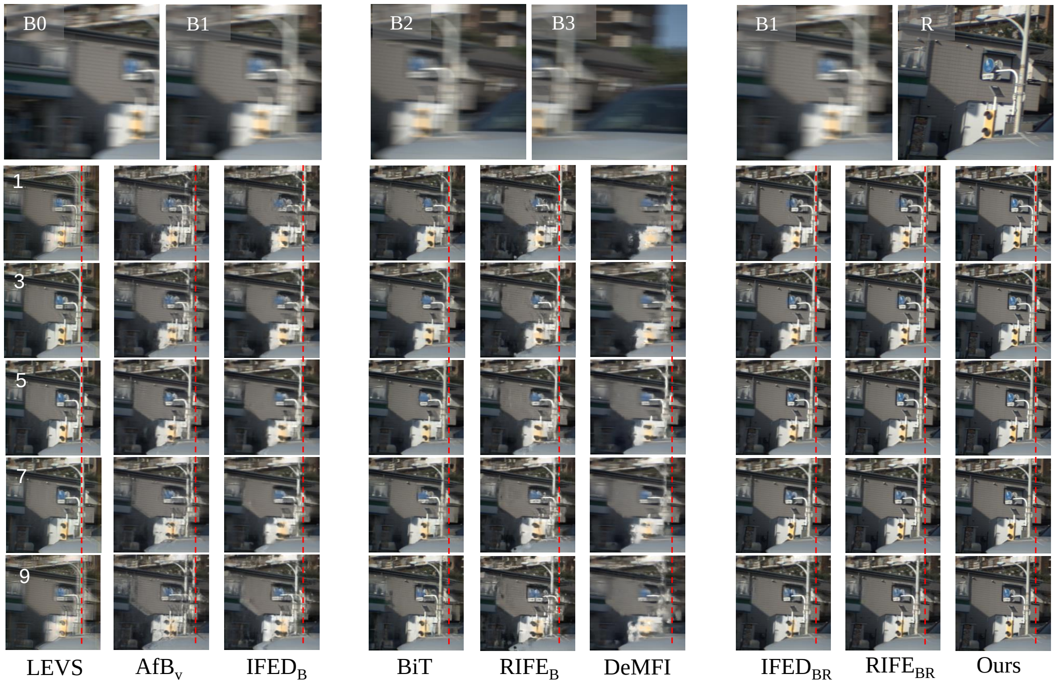
References

[1] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 2

[2] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video



Extended Figure A.7. **Visual results of comparisons with competitive settings** on synthetic data. I_t denotes the interpolated frame temporally located at time instant t.

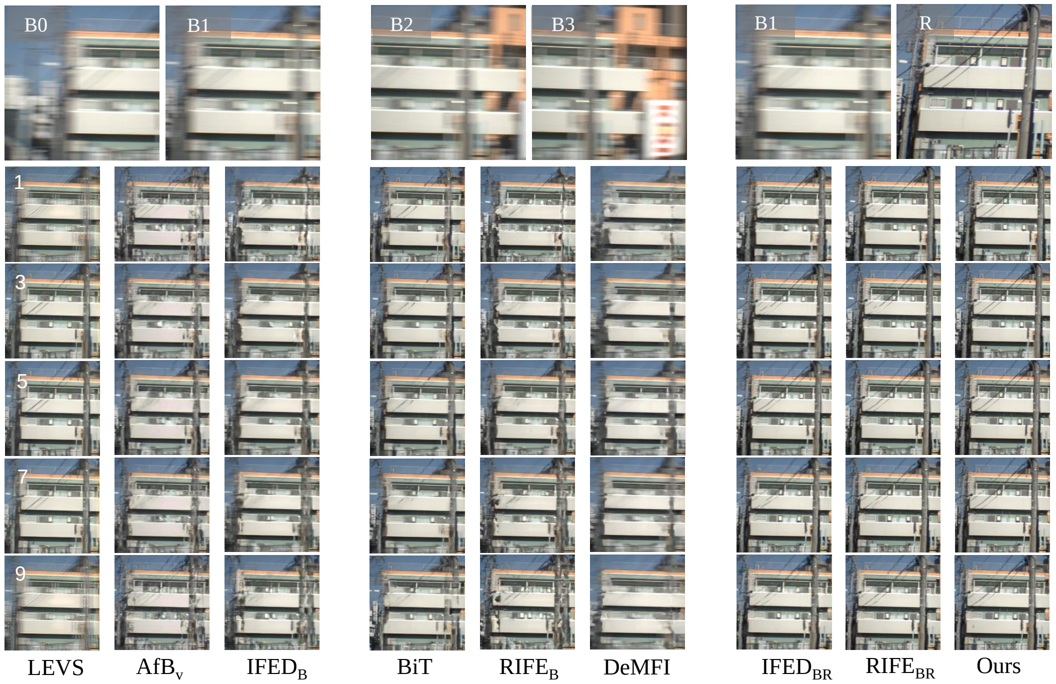


Extended Figure B.8. **Reconstructed consecutive frames from blurry inputs.** We present the multiple intermediate frames (5 out of 9) at different time generated by different models. The index k of each reconstrcuted frames denote their temporal location within exposure. B_0, B_1, B_2, B_3 are consecutive blur inputs and recovered latent frame corresponds to B_1. *Best viewed in zoom.*

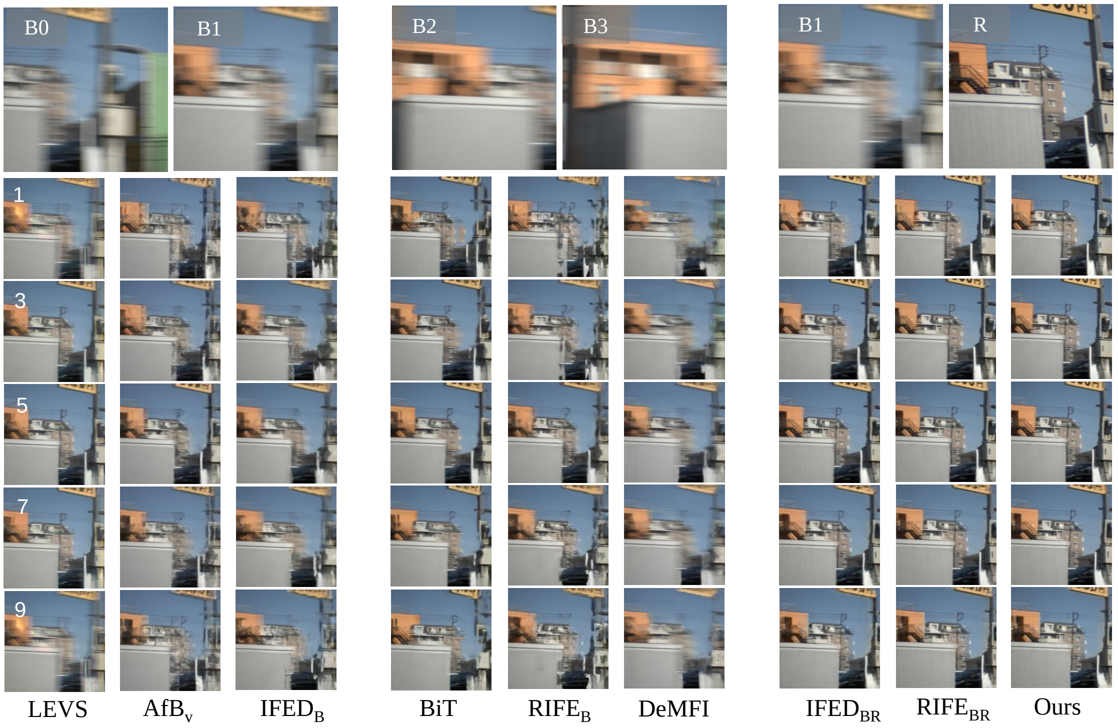
datasets for event cameras. in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3, 2020. 3

[3] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 624–642. Springer, 2022. 1, 3, 4, 7, 8

[4] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. In *Proceedings of the IEEE Conference on Computer Vision*



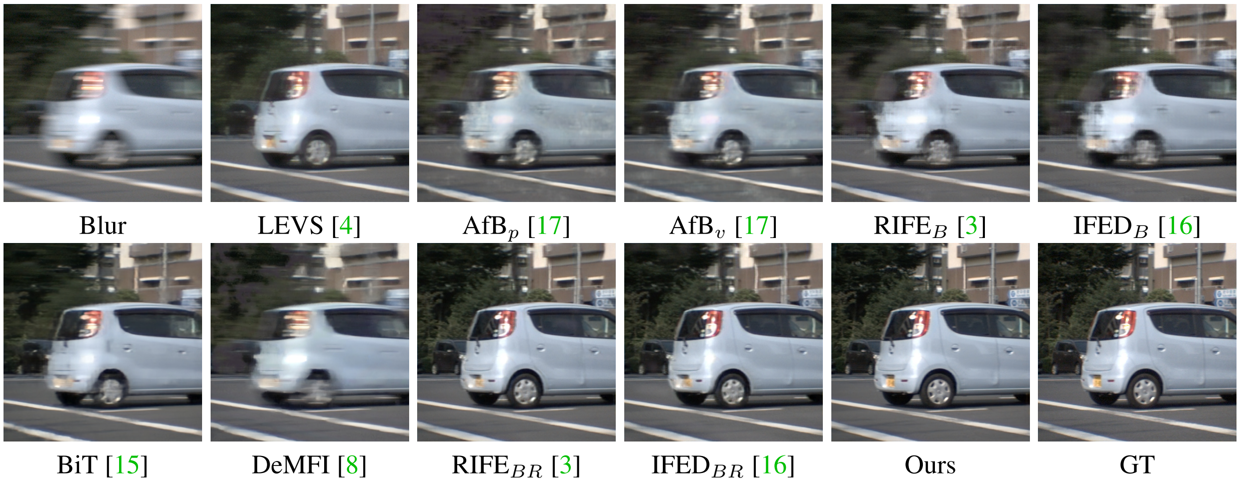
Extended Figure B.9. **Additional visual comparisons on video reconstruction of RealBR.** We present the multiple intermediate frames (5 out of 9) at different time generated by different models. The index k of each reconstrcuted frames denote their temporal location within exposure. B_0, B_1, B_2, B_3 are consecutive blur inputs and recovered latent frame corresponds to B_1. *Best viewed in zoom.*



Extended Figure B.10. **Additional visual comparisons on video reconstruction of RealBR.** We present the multiple intermediate frames (5 out of 9) at different time generated by different models. The index k of each reconstrcuted frames denote their temporal location within exposure. B_0, B_1, B_2, B_3 are consecutive blur inputs and recovered latent frame corresponds to B_1. *Best viewed in zoom.*

and Pattern Recognition, pages 6334–6342, 2018. 4, 7, 8

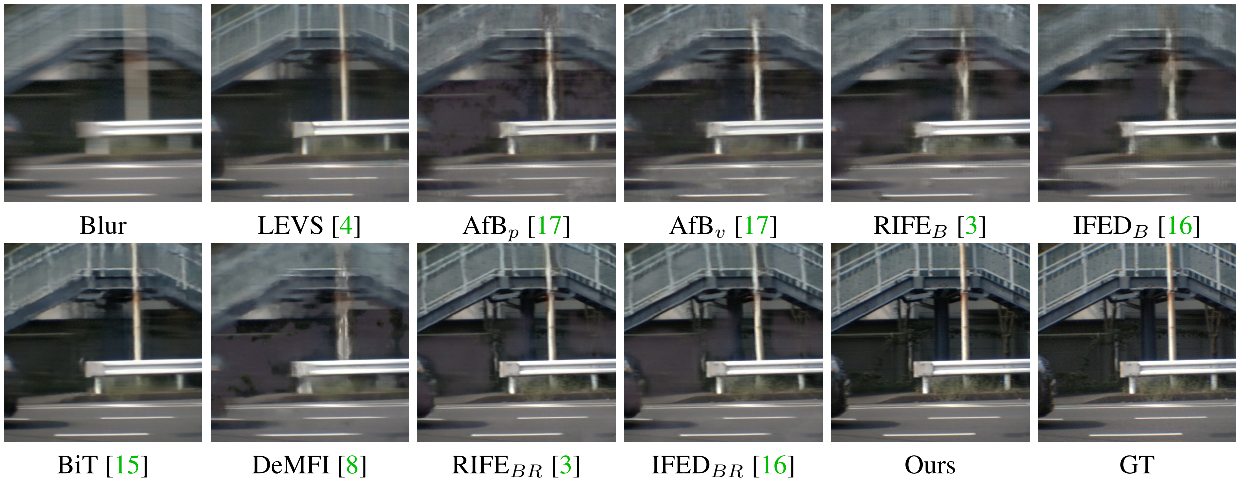
[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3



Blur LEVS [4] AfB$_p$ [17] AfB$_v$ [17] RIFE$_B$ [3] IFED$_B$ [16]

BiT [15] DeMFI [8] RIFE$_{BR}$ [3] IFED$_{BR}$ [16] Ours GT

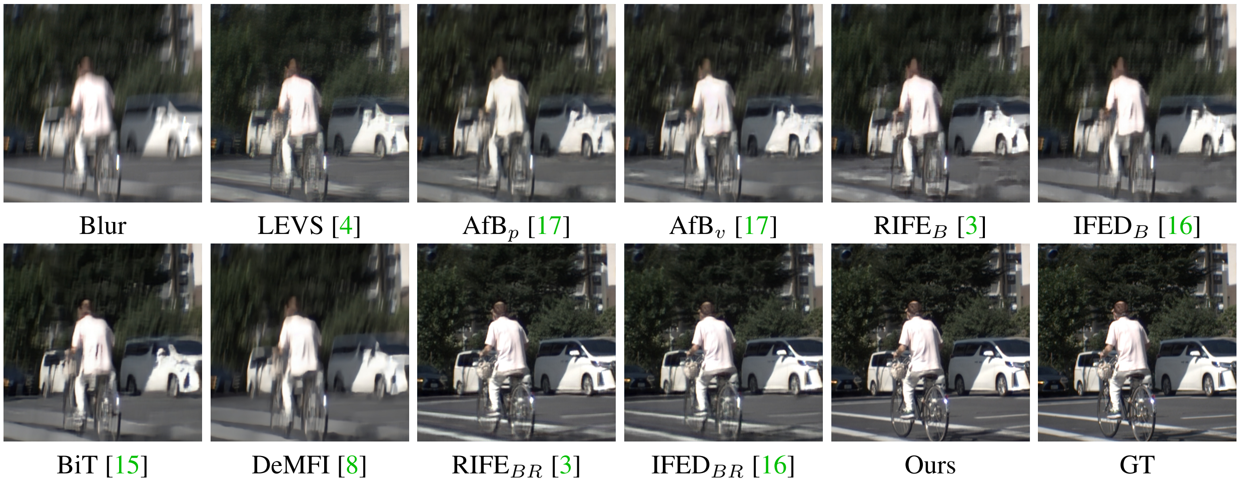
Extended Figure B.11. **Additional qualitative comparisons on RealBR.**



Blur LEVS [4] AfB$_p$ [17] AfB$_v$ [17] RIFE$_B$ [3] IFED$_B$ [16]

BiT [15] DeMFI [8] RIFE$_{BR}$ [3] IFED$_{BR}$ [16] Ours GT

Extended Figure B.12. **Additional qualitative comparisons on RealBR.**

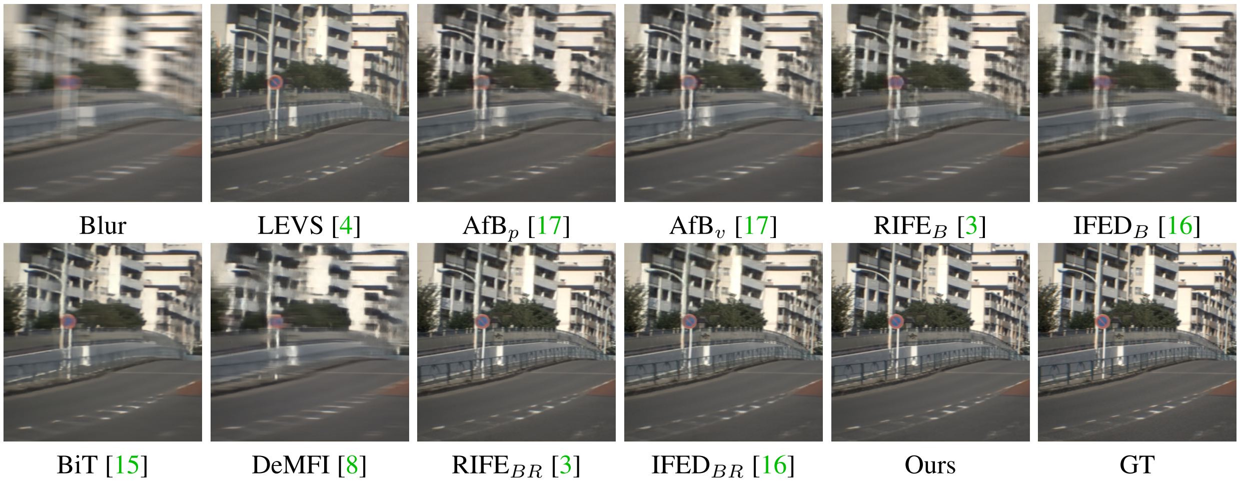


Blur LEVS [4] AfB$_p$ [17] AfB$_v$ [17] RIFE$_B$ [3] IFED$_B$ [16]

BiT [15] DeMFI [8] RIFE$_{BR}$ [3] IFED$_{BR}$ [16] Ours GT

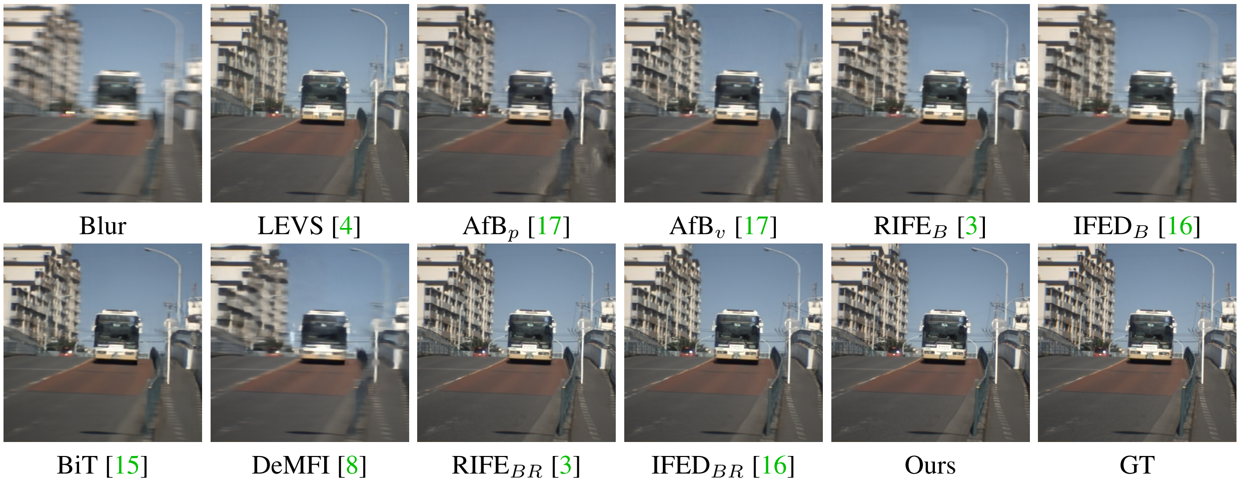
Extended Figure B.13. **Additional qualitative comparisons on RealBR.**

[6] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5949, 2020. 3


Blur | LEVS [4] | AfB$_p$ [17] | AfB$_v$ [17] | RIFE$_B$ [3] | IFED$_B$ [16]

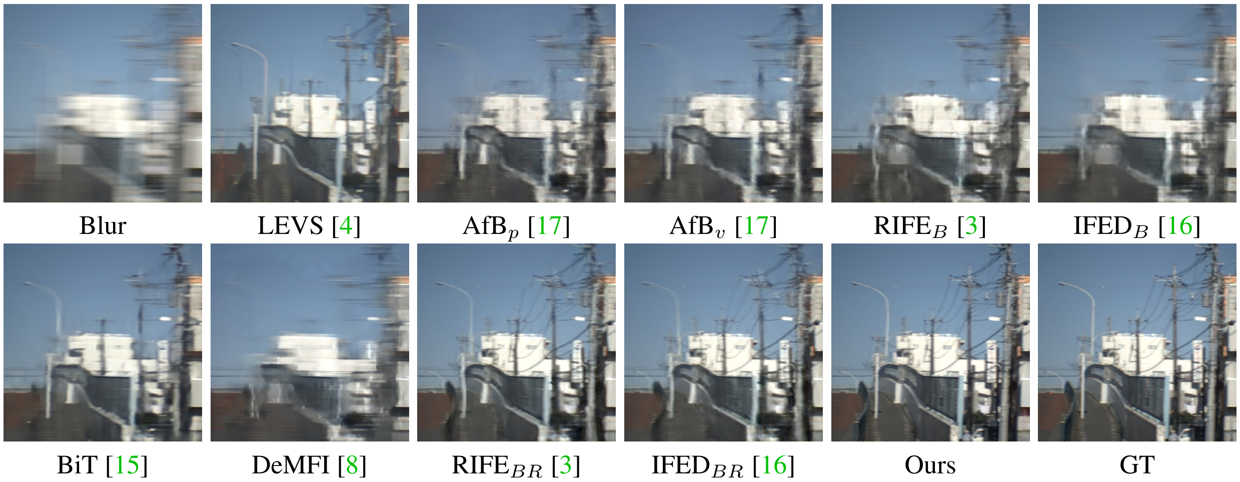
BiT [15] | DeMFI [8] | RIFE$_{BR}$ [3] | IFED$_{BR}$ [16] | Ours | GT

Extended Figure B.14. **Additional qualitative comparisons on RealBR.**


Blur | LEVS [4] | AfB$_p$ [17] | AfB$_v$ [17] | RIFE$_B$ [3] | IFED$_B$ [16]

BiT [15] | DeMFI [8] | RIFE$_{BR}$ [3] | IFED$_{BR}$ [16] | Ours | GT

Extended Figure B.15. **Additional qualitative comparisons on RealBR.**


Blur | LEVS [4] | AfB$_p$ [17] | AfB$_v$ [17] | RIFE$_B$ [3] | IFED$_B$ [16]

BiT [15] | DeMFI [8] | RIFE$_{BR}$ [3] | IFED$_{BR}$ [16] | Ours | GT

Extended Figure B.16. **Additional qualitative comparisons on RealBR.**

[7] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

2, 3

[8] Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 198–215. Springer, 2022. 2, 4, 7, 8

[9] Vijay Rengarajan, Shuo Zhao, Ruiwen Zhen, John Glotzbach, Hamid Sheikh, and Aswin C Sankaranarayanan. Photosequencing of motion blur using short and long exposures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 510–511, 2020. 3

[10] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 184–201. Springer, 2020. 1

[11] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 2, 3

[12] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based blurry frame interpolation under blind exposure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1588–1598, 2023. 3

[13] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020. 3

[14] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9219–9228, 2021. 1

[15] Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. *arXiv preprint arXiv:2211.11423*, 2022. 4, 7, 8

[16] Zhihang Zhong, Mingdeng Cao, Xiao Sun, Zhirong Wu, Zhongyi Zhou, Yinqiang Zheng, Stephen Lin, and Imari Sato. Bringing rolling shutter images alive with dual reversed distortion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 233–249. Springer, 2022. 1, 3, 4, 7, 8

[17] Zhihang Zhong, Xiao Sun, Zhirong Wu, Yinqiang Zheng, Stephen Lin, and Imari Sato. Animation from blur: Multimodal blur decomposition with motion guidance. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 599–615. Springer, 2022. 4, 7, 8

[18] Xinyu Zhou, Peiqi Duan, Yi Ma, and Boxin Shi. Evunroll: Neuromorphic events based rolling shutter image correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17775–17784, 2022. 3