
EPA: Boosting Event-based Video Frame Interpolation with Perceptually Aligned Learning - Supplementary Material -

Due to the limited space in the main text, we provide more details in the Supplementary Material. This supplementary material is comprised of the following sections.

Sec. 1 includes descriptions of different datasets used for model evaluation.

Sec. 2 conducts supplementary experiments both in qualitative and quantitative views.

Sec. 3 details the training and inference efficiency of our method.

Sec. 4 verified the effectiveness of our method in other degeneration scenarios.

Sec. 5 further verifies the core assumptions of our method.

Sec. 6 discusses the limitations of our method.

1 Datasets Details

1.1 Synthetic Event Datasets

To evaluate model performance in controlled environments, we employ several synthetic event datasets. These datasets typically generate event representations indirectly from high-frame-rate videos.

Vimeo90k-Triplet[5]: This dataset includes 3,782 image triplets in the test split, each at a resolution of 448×256 pixels. It contains diverse natural scenes, with moderate exercise, but with a smaller resolution.

GOPRO[2]: The test split consists of 11 scenes, each with an average of 1,100 images, captured at a resolution of 1280×720 pixels. It has a high frame rate but possesses a blur characteristic.

1.2 Real Event Datasets

To further evaluate the generalization of our model on real-world data, we utilize several real event datasets. These datasets are collected using actual event cameras (e.g., DAVIS, Prophesee) alongside conventional RGB cameras. They include realistic challenges such as occlusion, noise, and non-rigid motion, providing a more comprehensive assessment of model performance in practical settings.

High Speed Event and RGB Camera (HS-ERGB)[3]: The HS-ERGB dataset consists of 15 test scenes with RGB frames at resolutions exceeding 900×800 pixels. It is divided into two categories based on shooting distance—close and far—and includes challenging scenarios with occlusions and non-rigid motion, which are crucial for assessing model adaptability.

Beam Splitter Events & RGB (BS-ERGB)[4]: This large-scale dataset comprises 45 scenes in the training set, 17 in the validation set, and 26 in the testing set. The RGB images have a resolution of 970×625 pixels. It features substantial noise and complex non-rigid motion, serving as a strong benchmark for evaluating model robustness under real-world conditions.

EventAid-F[1]: It contains 10 test scenes with RGB images at resolutions greater than 954×636 pixels. It covers a wide range of motion patterns, making it particularly suitable for testing models under diverse and dynamic real-world motion conditions.

2 Additional Experiments

In this section, we present the complete evaluation results on the EventAid-F dataset. All models participating in the evaluation use weights trained on the BS-ERGB dataset, which is characterized by extensive noise. As shown in Tab. 1, our method demonstrates superior generalization performance, achieving comparable results on conventional metrics such as PSNR and SSIM, while significantly outperforming others on perceptual quality metrics. Notably, our method achieves a 45% improvement on the LPIPS perceptual metric compared to the second-best result, which validates the effectiveness and feasibility of our approach.

Table 1: Performance comparison on EventAid-F datasets. The best results are marked in **Bold** while the second ones are marked with underlines.

Method	EventAid-F									
	7 skip					15 skip				
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FloLPIPS \downarrow	DISTS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FloLPIPS \downarrow	DISTS \downarrow
Timelens	34.285	0.939	<u>0.031</u>	0.068	0.044	31.800	<u>0.910</u>	0.051	0.108	0.064
CBMNet	33.318	0.930	0.044	0.088	0.053	32.273	0.921	0.055	0.113	0.061
TLXNet	32.361	0.910	0.033	<u>0.067</u>	0.041	30.096	0.876	0.046	0.092	0.052
EPA (ours)	<u>33.906</u>	<u>0.932</u>	0.017	0.055	<u>0.043</u>	<u>31.987</u>	<u>0.910</u>	0.024	0.076	<u>0.055</u>

2.1 More Visual Results on BS-ERGB dataset

In the Figs. 3 and 4, we show additional qualitative results on the BS-ERGB dataset.

2.2 More Visual Results on HS-ERGB dataset

In the Figs. 5 to 7, we show additional qualitative results on the HS-ERGB dataset.

2.3 More Visual Results on EventAid-F dataset

In the Figs. 8 and 9, we show additional qualitative results on the EventAid-F dataset.

3 Efficiency Analysis

3.1 Training cost

All experiments were conducted on a single NVIDIA RTX 3090 GPU. Our training process is composed of two distinct stages:

Stage 1 (Reconstruction Generator Training): The model was trained for 100 epochs using a single RGB image with a batch size of 28 and an image size of 256×256 , requiring approximately 24 hours.

Stage 2 (BEGA Module Training): Following the first stage, the weights of the other parts were frozen, and only the BEGA module was trained for 40 epochs. This stage utilized two RGB images and the corresponding event streams, with a batch size of 32. This stage took approximately 48 hours to complete.

3.2 Inference Cost and Comparison

To evaluate the efficiency of our model, we compare its resource consumption and runtime against other state-of-the-art methods. As detailed in Table 2, our approach achieves a highly competitive runtime and model size, demonstrating an excellent balance between performance and efficiency.

Table 2: Comparison of model size, computational cost, and performance. \dagger represents trainable parameters. PSNR and LPIPS scores come from the real dataset HS-ERGB.

Method	Parameters \dagger	FLOPs	Runtime	PSNR \uparrow	LPIPS \downarrow
Timelens	79.2M	18.7B	0.033s	31.871	0.053
CBMNet	22.2M	46.7B	0.231s	31.876	0.101
TLXNet	7.87M	16.5B	0.027s	31.578	0.046
EPA (ours)	13.9M	20.2B	0.086s	33.402	0.015

4 Robustness to Other Degradation Types

To address the concern that our method’s robustness was primarily demonstrated on motion blur and camera noise, we conducted additional experiments to evaluate its performance against other common image degradation types. Specifically, we tested all competing methods under two challenging conditions: JPEG compression artifacts and low-resolution degradation.

The quantitative results are presented in Table 3. As shown, our method consistently achieves the best perceptual quality (LPIPS, FloLPIPS, DISTS) under both degradation settings, confirming its strong robustness. Although a traditional method like CBMNet may achieve a slightly higher PSNR score, it suffers a significant collapse in perceptual quality, often introducing severe color artifacts in moving objects. In contrast, TLXNet, being purely reliant on optical flow, fully propagates the degradation from the keyframes, resulting in large discrepancies with the non-degraded ground truth and severely hampering its performance, as shown in Fig. 1.

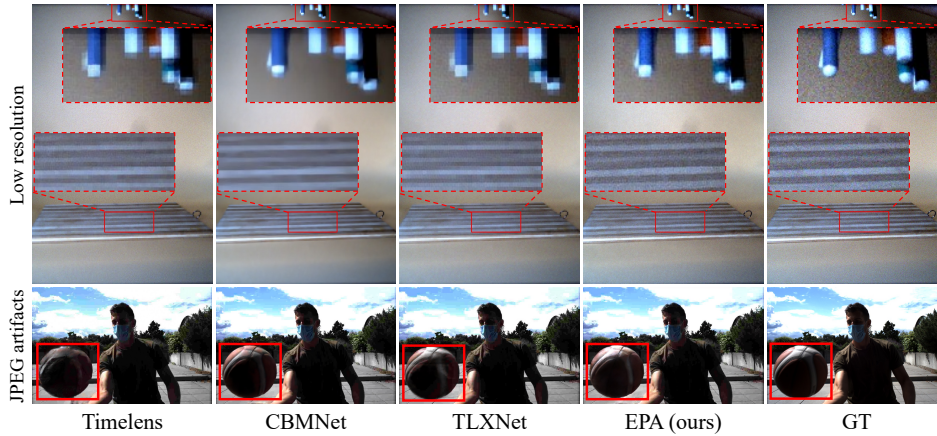


Figure 1: Comparison of different methods under different degradation scenarios visualization.

These results further validate our central claim: by operating in a degradation-insensitive semantic-perceptual feature space, our method maintains superior performance across a wider range of challenging conditions.

Table 3: Quantitative comparison on the BS-ERGB pen_03 scene under JPEG compression and low-resolution degradation. Our method demonstrates superior perceptual quality in both scenarios.

Degradation Type	Method	PSNR \uparrow	LPIPS \downarrow	FloLPIPS \downarrow	DISTS \downarrow
JPEG Compression	Timelens	21.373	0.216	0.313	0.235
	CBMNet	22.082	0.575	0.818	0.277
	TLXNet	20.126	0.371	0.599	0.268
	Ours	21.423	0.141	0.246	0.177
Low Resolution	Timelens	21.285	0.191	0.253	0.247
	CBMNet	22.022	0.551	0.743	0.274
	TLXNet	20.175	0.226	0.382	0.247
	Ours	21.584	0.135	0.212	0.180

5 Quantitative Analysis of Feature Robustness

To provide a more rigorous analysis beyond the L2 distance comparison presented in the main paper, we quantitatively evaluate the degradation-insensitivity of the DINO features. We adopt the Signal-to-Noise Ratio (SNR) as a metric to compare the stability of features against the stability of the input images under various common degradations.

The results of this analysis are detailed in Table 4. We calculated the SNR for both the raw input images (Image level) and the DINO features extracted from different network depths (Feature levels 0, 1, and 2).

Table 4: Signal-to-Noise Ratio (SNR) comparison across different degradation types. Lower SNR values indicate greater stability and robustness to degradation. The DINO features, especially at deeper levels, exhibit significantly more stable SNR values than the raw image pixels.

Type	Gaussian blur	Noise	JPEG artifact	Low resolution	Motion blur
Image level	58.07	32.61	85.30	31.07	29.29
Feature level 0	9.20	2.22	14.75	4.67	4.44
Feature level 1	3.33	1.91	2.20	2.12	2.13
Feature level 2	2.41	1.22	1.26	1.53	1.85

The data clearly demonstrates that while the SNR at the image level fluctuates significantly with different degradations, the SNR within the DINO feature space is substantially more stable. Notably, this robustness increases in deeper, more semantic layers (Feature levels 1 and 2). This quantitatively validates our central hypothesis that the DINO feature space is inherently more robust to degradation, which is the cornerstone of our proposed feature-level supervision strategy.

6 Limitation

Specifically, although our method incorporates feature-level supervision and leverages event-based guidance to enhance temporal consistency and improve perceptual quality, the use of a generator-based architecture inevitably introduces a certain degree of randomness. Compared with optical flow-based approaches, our method may exhibit discrepancies in color fidelity relative to the ground truth. As illustrated in Fig. 2, such color inconsistencies become particularly noticeable in scenarios involving large motion. In future work, we will focus on enhancing the consistency of generative video interpolation models, with the goal of further improving structural accuracy and perceptual fidelity.

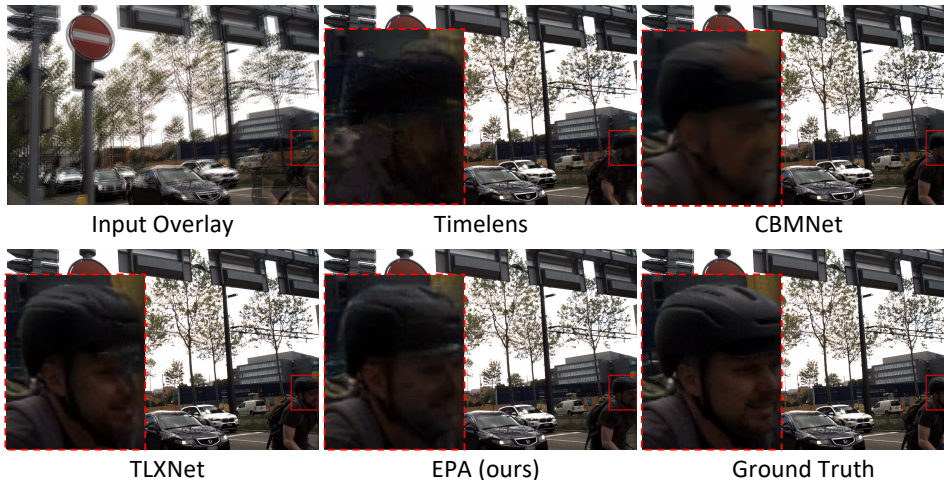


Figure 2: Visual results on the real dataset.

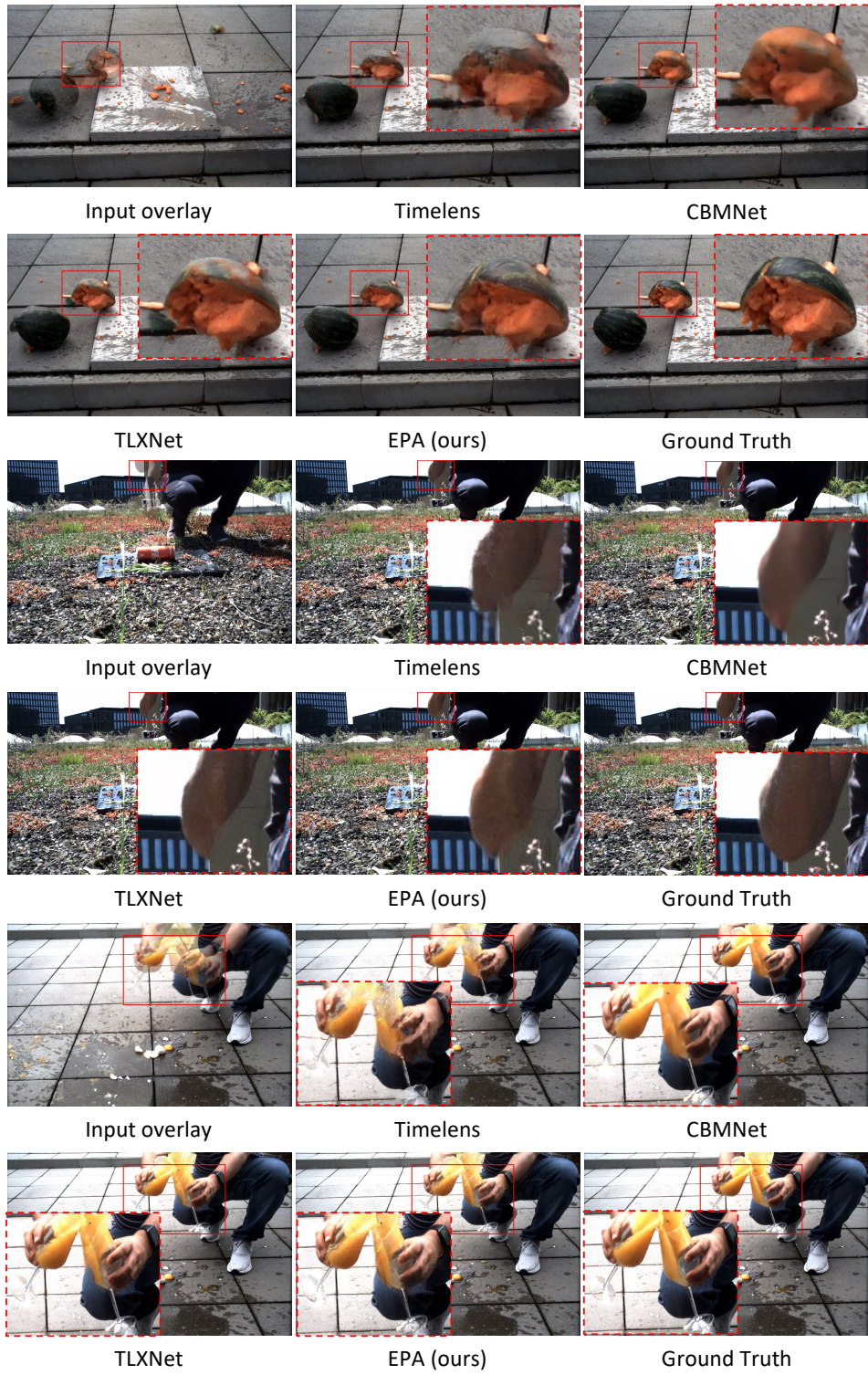


Figure 3: Visual results on the real dataset.

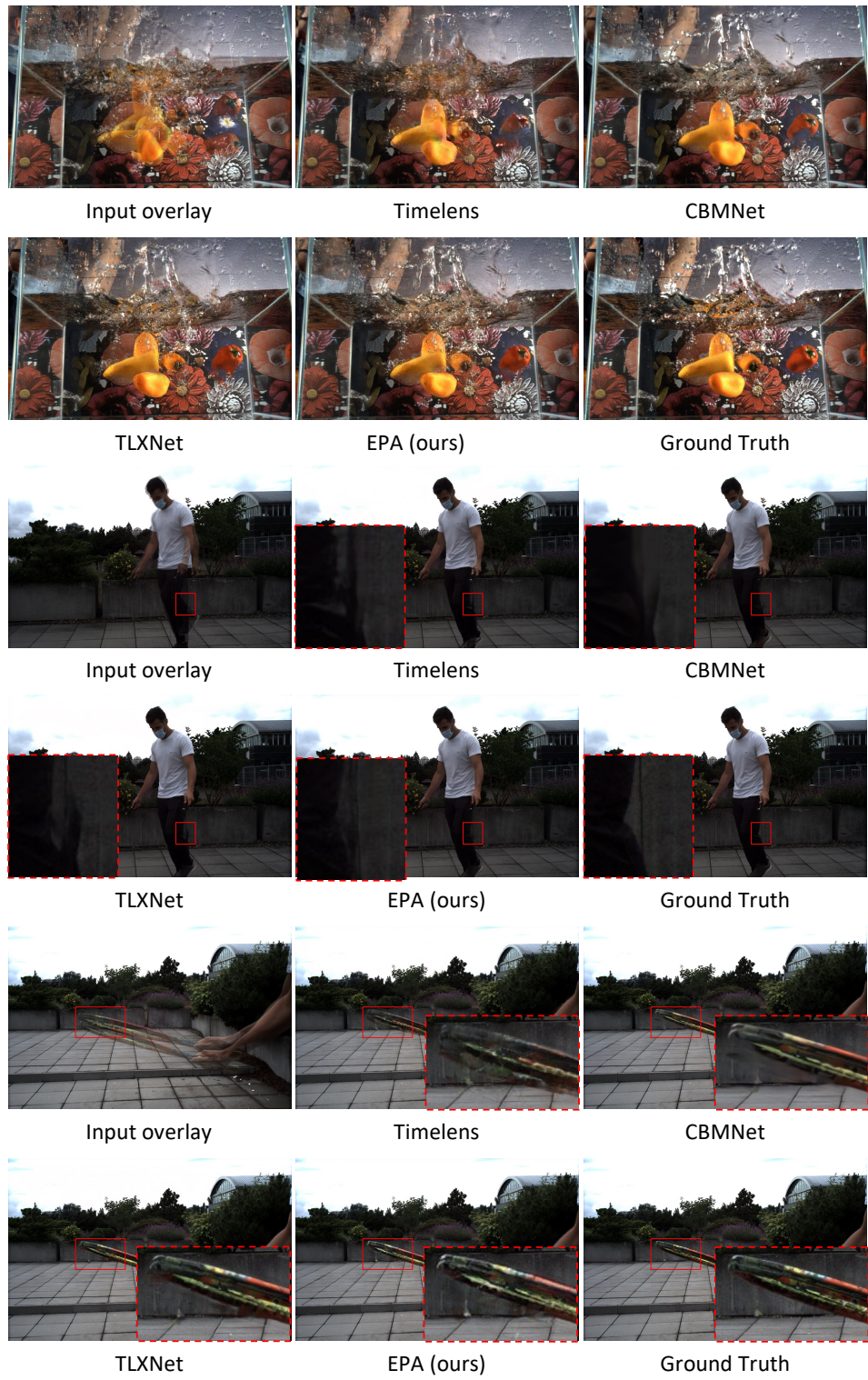


Figure 4: Visual results on the real dataset.

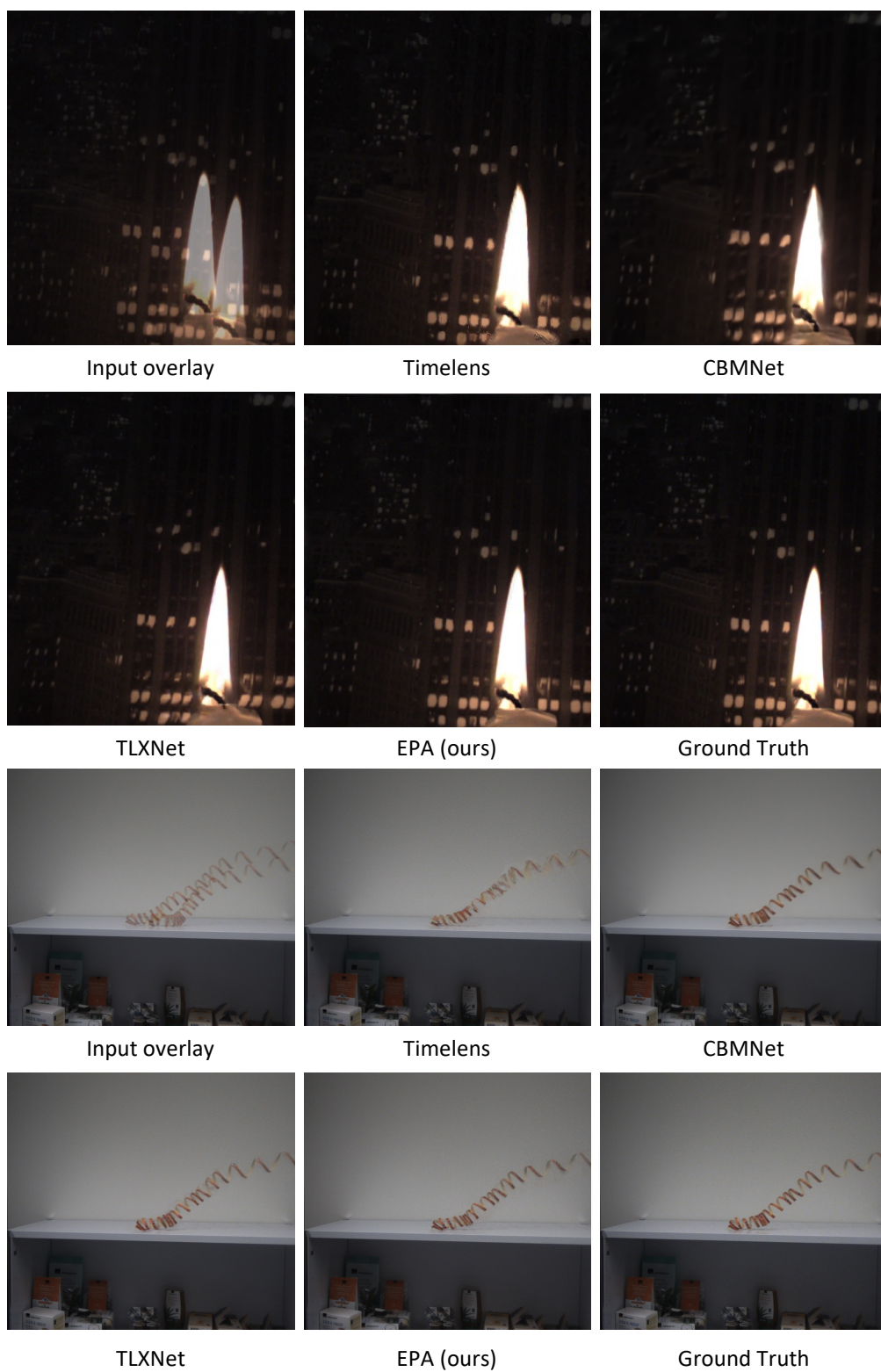


Figure 5: Visual results on the real dataset

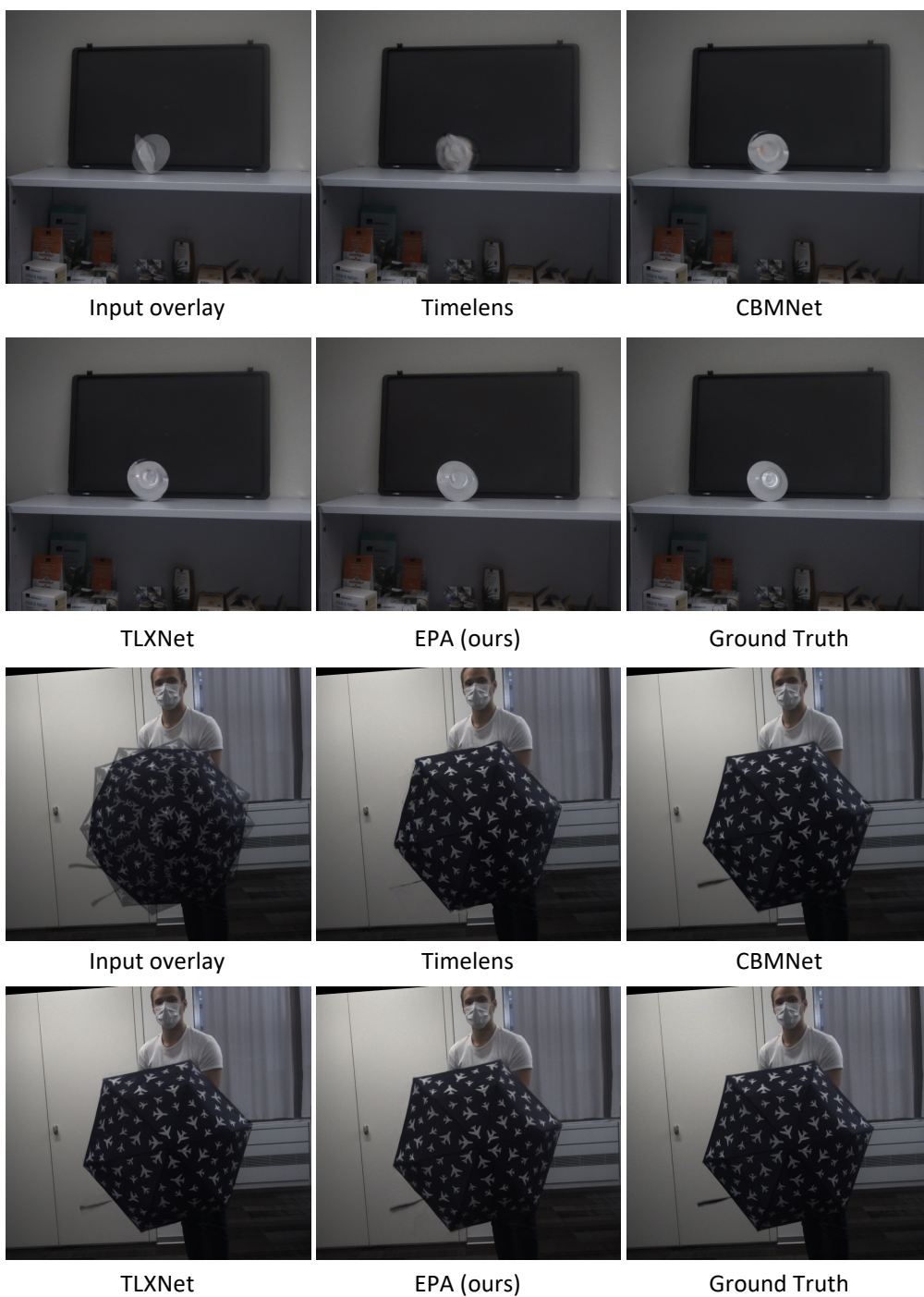


Figure 6: Visual results on the real dataset

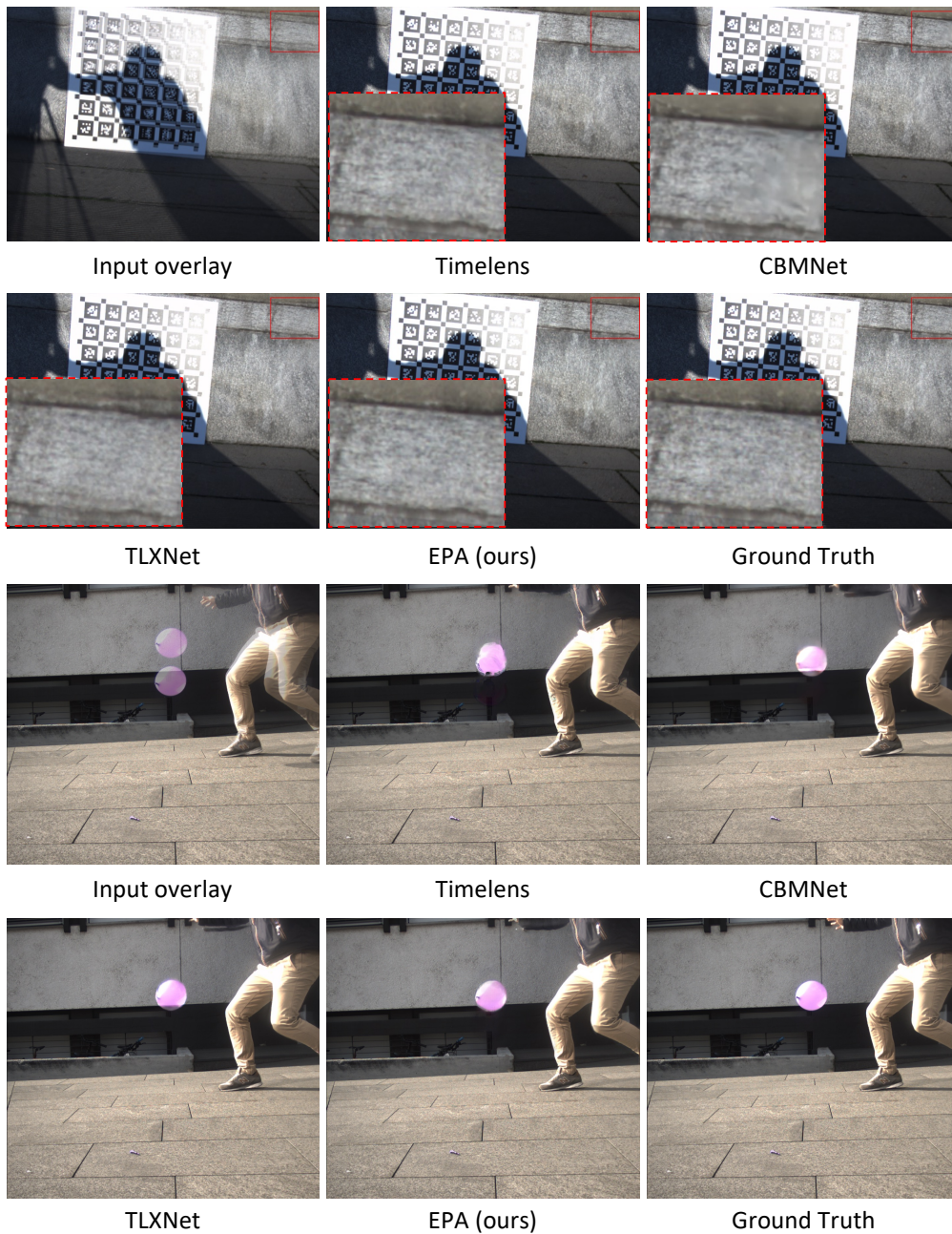


Figure 7: Visual results on the real dataset

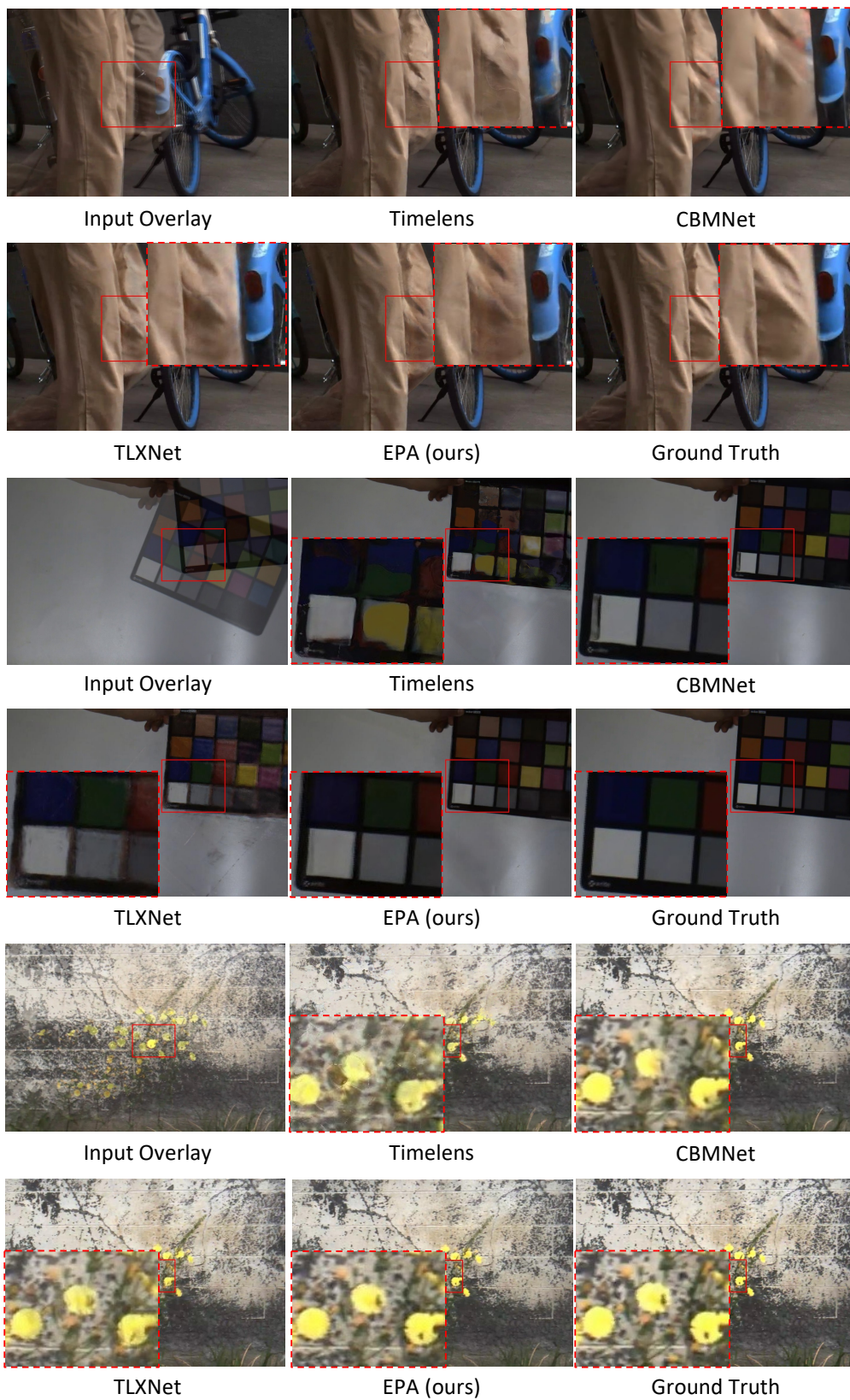


Figure 8: Visual results on the real dataset

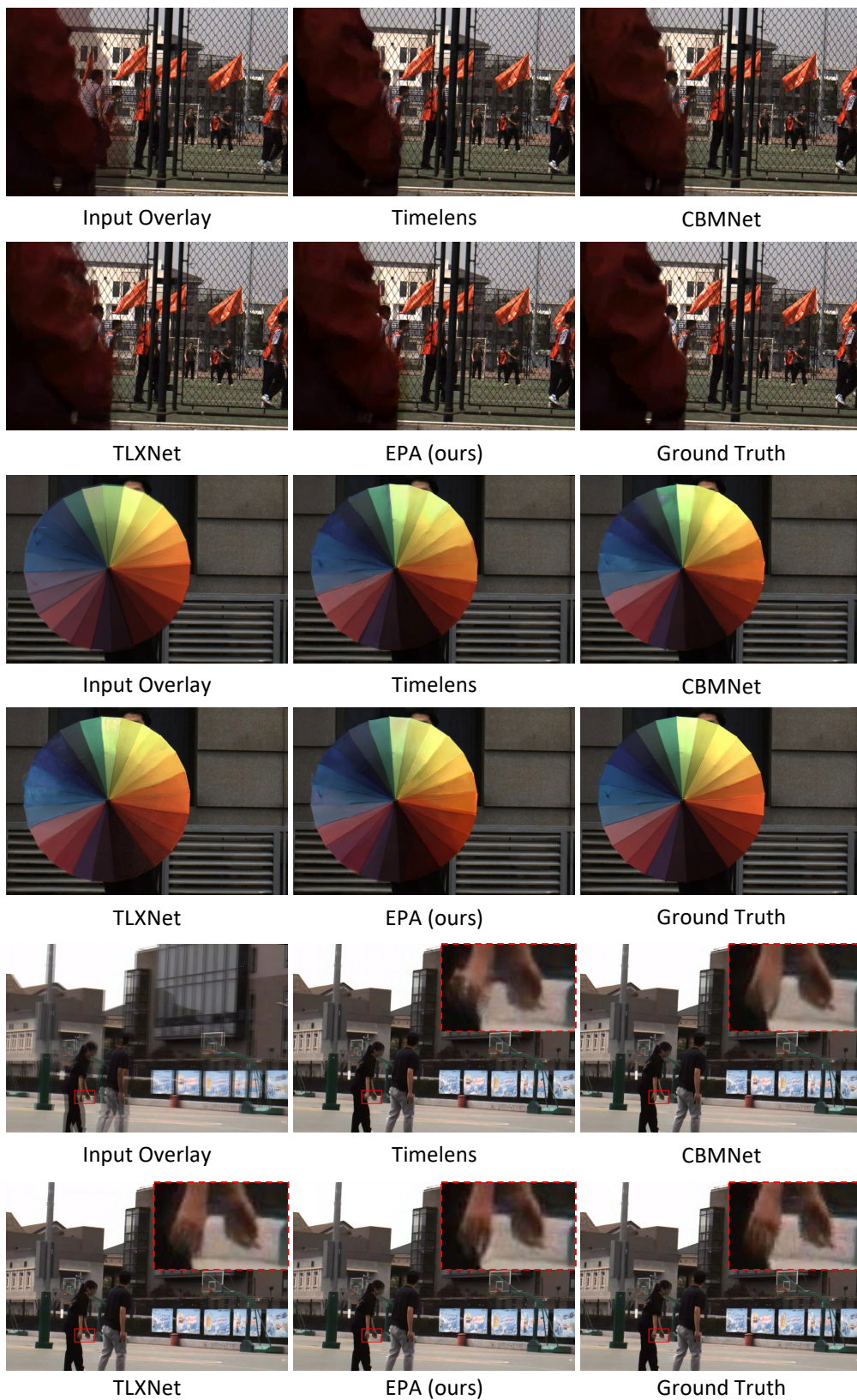


Figure 9: Visual results on the real dataset

References

- [1] Peiqi Duan, Boyu Li, Yixin Yang, Hanyue Lou, Minggui Teng, Yi Ma, and Boxin Shi. Eventaid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. *arXiv preprint arXiv:2312.08220*, 2023.
- [2] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.
- [3] 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.
- [4] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022.
- [5] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.