Coherent Event Guided Low-Light Video Enhancement (Supplementary Material)

Jinxiu Liang[1,2] Yixin Yang[1,2] Boyu Li[1,2] Peiqi Duan[1,2] Yong Xu[3] Boxin Shi*[1,2]

[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[3] School of Computer Science and Engineering, South China University of Technology

{cssherryliang, yangyixin93, liboyu, duanqi0001, shiboxin}@pku.edu.cn yxu@scut.edu.cn

Figure 6. Hybrid camera system.

6. Details of the Hybrid Camera System

We build a hybrid camera system to evaluate the proposed method for real-world scenarios. It consists of an event camera DAVIS346 and an industrial RGB camera FLIR Chameleon 3 Color, as shown in Figure 6. The two sensors are connected via a Thorlabs CCM1-BS013 beam splitter mounted in front of them. We perform a coarse geometric alignment between them by using a checkerboard. Due to the large modality gap and weakened features in low-light frames, such an alignment is usually not precise. This issue is further relieved by the proposed spatial coherence modeling module, which allows feature-level alignment between events and low-light frames.

Camera parameter requirements It depends on the motion and lighting of the scenes. We assume that the exposure time is chosen to mitigate any blurring effects while the ISO is set to prevent the average intensity from being too low. Detailed settings of visual results on real data are shown in Table 4.

7. More Implementation Details

Following common practice in flow-based video restoration [7, 11, 1, 22], the optical flow estimators $\mathcal{F}^E_{\text{flow}}$ and $\mathcal{F}^L_{\text{flow}}$ are initialized from pretrained models for events [4] and frames [14], respectively. In particular, the parameters of $\mathcal{F}^L_{\text{flow}}$ are pre-trained on FlyingChairs [2] and FlyingThings3D [10], which are officially provided by PyTorch. The parameters of $\mathcal{F}^E_{\text{flow}}$ are pretrained on DSEC [3].

Table 4. Camera parameter requirements of the visual results.

Setting	Fig. 1	Fig. 5	Fig. 7	Fig. 12	Fig. 13
Exposure time (ms)	10	0.3	0.63	0.15	20
ISO	200	200	200	200	200
Luminance (cd/m2)	1.4	160.7	234.9	385.0	0.6

To compensate for the resolution gap and sensor misalignment, the global coherence is estimated from multimodal coherence C^{modal} by $\mathcal{F}_{\text{global}}$ from a projection matrix initialized as an identity mapping. In particular, projected coordinates are used to extract the correlation slice from multimodal coherence C^{modal} similar to Eq. (8). The basic units of $\mathcal{F}_{\text{global}}$ include a 3×3 convolutional block, a group normalization + ReLU, and a max-pooling layer with stride 2. They are used to continuously downsample the input features until their spatial resolution reaches 2×2, which is then projected into a $2 \times 2 \times 2$ displacement cube $\boldsymbol{D}$ by a convolutional layer.

For more details, please refer to the code provided on our project page[1].

8. Additional Analysis

Comparison with stronger competitors. To demonstrate the effectiveness of the proposed multimodal coherence modeling module and the temporal coherence propagation module, we compare the proposed method with four stronger competitors built upon the state-of-the-art low-light video enhancement methods SDSD [17] and MBLLEN [8], event-guided frame interpolation method Time Lens [16], and a low-light events-to-video reconstruction method DVS-Dark [21].

[1] https://sherrycattt.github.io/EvLowLight

Table 5. Quantitative comparison with stronger competitors on our synthetic data with severe noise. ↑ (↓) means higher (lower) is better.

Method	PSNR↑	SSIM↑	LPIPS↓
MBLLEN† [8]	20.81	0.7928	0.3403
SDSD† [17]	21.20	0.7280	0.3962
Time Lens† [16]	18.60	0.7106	0.3699
DVS-Dark† [21]	6.25	0.0507	0.8131
Ours	**23.98**	**0.8369**	**0.2794**

Table 6. Comparison of computational complexity. ↑ (↓) means higher (lower) is better. The values of PSNR are also provided for reference.

Method		PSNR↑	FLOPs↓	#Params↓
Pure event	DVS-Dark [21]	9.81	721.27 G	11.89 M
	E2VID [12]	16.20	139.62 G	10.71 M
Image-based	SCI [9]	15.96	0.06 G	0.0003 M
	Transformer [19]	15.81	112.30 G	39.12 M
	URetinex-Net [18]	20.87	266.88 G	0.34 M
Video-based	MBLLEN [8]	17.77	210.34 G	0.12 M
	SDSD [17]	12.60	214.44 G	4.30 M
	StableLLVE [20]	19.37	47.21 G	4.32 M
Hybrid	Ours	23.98	175.39 G	15.03 M

The inputs of SDSD [17] and MBLLEN [8] are multiple frames that can benefit from interframe information provided by an event camera. In particular, events are provided as additional input through the additional sixth convolutional layers, which share the architecture similar to [13]. The generated features of events are resized and concatenated with the input frames, which are then fused by a 3×1 convolutional layer and fed to the main model of SDSD [17] and MBLLEN [8]. All the layers are then trained end-to-end following their recommended settings. Time Lens [16] accepts hybrid inputs of events and two adjacent frames of the target to be predicted. We adjust its network architecture to take in a low-light version of the target frame and its adjacent low-light frames. DVS-Dark [21] are initially designed for events-to-video reconstruction in the dark, relying solely on event inputs. Its network architecture is modified to accept additional frame inputs.

Quantitative results are shown in Table 5, where the compared methods are denoted as SDSD†, MBLLEN†, Time lens†, and DVS-Dark†, respectively. With additional inputs of events, SDSD† and MBLLEN† perform better than their vanilla version. The modified versions of Time Lens [16] and DVS-Dark† to accept additional inputs of the corresponding low-light versions of the target frame show about 5dB and 17dB drops compared to the proposed method in terms of PSNR, due to inadequate noise-handling mechanisms in low-light conditions. The proposed method performs the best among all compared methods.

Computational cost. To compare the computational complexity of some recently proposed deep learning-based methods and our method, we report floating point operations (denoted by FLOPs) and the number of parameters (denoted by #Params) in Table 6. FLOPs are computed by averaging each method's processing for 100 frames at a resolution of 640×480. The PSNR values are also provided for reference. It can be seen that our method performs the best with moderate cost.

The proposed components – global feature alignment, pixel-wise motion aggregation, temporal coherence propagation, and exposure parameter estimation – constitute only 5.19%, 8.05%, 2.86%, and fewer than 0.01% of the total parameters, respectively.

Robustness to sensor misalignment. To compensate for the low resolution of events and let them match their frames counterparts, hybrid frame-event camera systems are often adopted in event guided image/video enhancement tasks [6, 16, 15]. They could be optically colocated via a beam splitter [6, 15] or put as close as possible side by side [16]. As mentioned in the main text, aligning the two sensors precisely is difficult. For example, the events and frames shown in the first row of Figure 7 (c). When their estimated homography parameters are directly applied to data that capture a scene with different depths, there is still misalignment between events and frames, as shown in the third row of Figure 7 (c). Online registration for each scene (with different depths and lighting variations) is necessary to obtain stable results for hybrid camera systems [16, 15]. However, it becomes fragile under low light conditions, since the features for computing homography are too weak to be precisely extracted. Thanks to the proposed multimodal coherence modeling module, the proposed method is robust to misalignment between events and frames. As shown in Figure 7 (d), the proposed method stably produces good denoising results under aligned (the first row), misaligned (the third row), and unaligned (the second and the fourth row) settings.

Robustness to discontinuity of motion and lighting inconsistency. The assumption of brightness constancy and continuity of motion limits the performance of frame-based optical flow. In this paper, we introduce the event camera for capturing inter-frame motion in the order of microseconds and brightness changes in a high dynamic range over 120 dB. For example, the cat's head presented in Figure 8 (a) becomes occluded in Figure 8 (c), whose motion between frames is discontinuous. The cat basking in the sun in Figure 8 (a) becomes covered in the girl's shadow in Figure 8 (c), whose brightnesses are inconsistent. Nevertheless, the high temporal resolution and high dynamic range of events essentially increase the robustness to deal with the discontinuity of motion ((b) the cat's head occluded) and the lighting inconsistency ((c) the cat in shadow).

9. More Visual Comparison Results

We provide more visual comparisons on both synthetic data and real data. The compared methods include *(i) image-based methods*: LIME [5], SCI [9], Transformer [19], URetinex-Net [18]; *(ii) video-based methods*: MBLLEN [8], StableLLVE [20], and SDSD [17]; *(iii) event-based restoration methods*: DVS-Dark [21] and E2VID [12]; *(iv) methods with hybrid inputs of events and frames*: the aforementioned stronger competitors built upon methods with multiple input frames, *i.e.*, MBLLEN† [8] and SDSD† [17], and the proposed method. The results of all existing methods are produced by their officially released codes with recommended parameter settings. LIME [5] is the state-of-the-art conventional method, while the others are all learning-based methods. Visual comparison results on synthetic data are shown in Figure 9, Figure 10, and Figure 11, while results on real data are shown in Figure 12 and Figure 13.

Please refer to the *video* on our project page[2] for low-light video enhancement results of different methods on both synthetic and real data. The video results with more recovered details and reduced noise compared to the other methods demonstrate the effectiveness of the proposed temporal propagation module for capturing redundancy between consecutive frames.

DSEC [3] is a stereo event camera dataset for driving scenarios containing several real samples of paired events and frames captured at night. As shown in the first four rows of Figure 14, the proposed method can recover details of buildings and license plate numbers of fast-moving cars, demonstrating its robustness in real complex scenarios.

Time Lens [16] proposes an event-guided frame interpolation method and provides the corresponding real data pairs of events and frames. Frames are captured with a fast shutter speed, which makes them contain noise. To demonstrate the robustness of the proposed method to normal-light images, we provide the corresponding results on real data provided in Time Lens [16] in the last two rows of Figure 14.

10. Qualitative Results of Ablation Studies

We compare the proposed method with its five variants to validate the effectiveness and necessity of each component: *(i)* w/o events: without event as additional inputs; *(ii)* w/o $\mathcal{F}_{\text{spat}}$: without the proposed spatial coherence modeling module $\mathcal{F}_{\text{spat}}$; *(iii)* w/o feature alignment: without the proposed feature-level alignment described in Eq. (13) of the main text; *(iv)* w/o $\mathcal{F}_{\text{temp}}$: without the proposed temporal coherence propagation module $\mathcal{F}_{\text{temp}}$; *(vi)* w/o noise simulation: without the proposed noise simulation process that considers complex degradation in real-world scenarios.

We provide visual results in Figure 15 and Figure 16 corresponding to the quantitative results shown in Table 3 of the main text. We introduce the event camera, which can extract precise temporal information but has quite a different representation of visual scenes, to better utilize temporal redundancy for video restoration and enhancement. In low-light frames, features are weakened, and motion information is hard to be estimated. As shown in Figure 15 (d), (e), and (f), without the help of events or sophisticated alignment schemes between the two sensors, oversmooth results are produced, *e.g.*, the textures of the tire disappear as shown in the blue boxes in Figure 15 (d), (e), and (f). The proposed temporal coherence propagation module $\mathcal{F}_{\text{temp}}$ is used to better extract and propagate temporal redundancy information over time with the help of the event camera. Without it, the noise reduction performance decreases, as shown in the red boxes in Figure 16 (g). To develop a low-light video enhancement method with better generalization ability in complex real-world scenarios, we propose to use a more practical degradation process for synthesizing data for training. Its effectiveness is validated in both Table 3 of the main text and visual results in Figure 16 (h).

[2] https://sherrycattt.github.io/EvLowLight

References

[1] Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 1

[2] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v d Smagt, D. Cremers, and T. Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *Proc. of International Conference on Computer Vision*, 2015. 1

[3] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A Stereo Event Camera Dataset for Driving Scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, July 2021. 1, 3, 12

[4] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense Optical Flow from Event Cameras. In *Proc. 3DV*, 2021. 1

[5] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, Feb. 2017. 3, 7, 8, 9, 10, 11

[6] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic Camera Guided High Dynamic Range Imaging. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 2

[7] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Recurrent Back-Projection Network for Video Super-Resolution. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1

[8] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: Low-Light Image/Video Enhancement using CNNs. In *Proc. of British Machine Vision Conference*, 2018. 1, 2, 3, 7, 8, 9, 10, 11

[9] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward Fast, Flexible, and Robust Low-Light Im-

age Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2, 3, 7, 8, 9, 10, 11

[10] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 1

[11] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded Deep Video Deblurring Using Temporal Sharpness Prior. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 1

[12] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-Video: Bringing Modern Computer Vision to Event Cameras. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2, 3, 7, 8, 9, 10, 11

[13] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert E. Mahony, and Davide Scaramuzza. Fast Image Reconstruction with an Event Camera. In *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020. 2

[14] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Proc. of European Conference on Computer Vision*, 2020. 1

[15] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Geourgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric nonlinear flow and multi-scale fusion. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2

[16] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time Lens: Event-based video frame interpolation. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 1, 2, 3, 12

[17] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing Dynamic Scene in the Dark: A High-Quality Video Dataset with Mechatronic Alignment. In *Proc. of International Conference on Computer Vision*, 2021. 1, 2, 3, 7, 8, 9, 10, 11

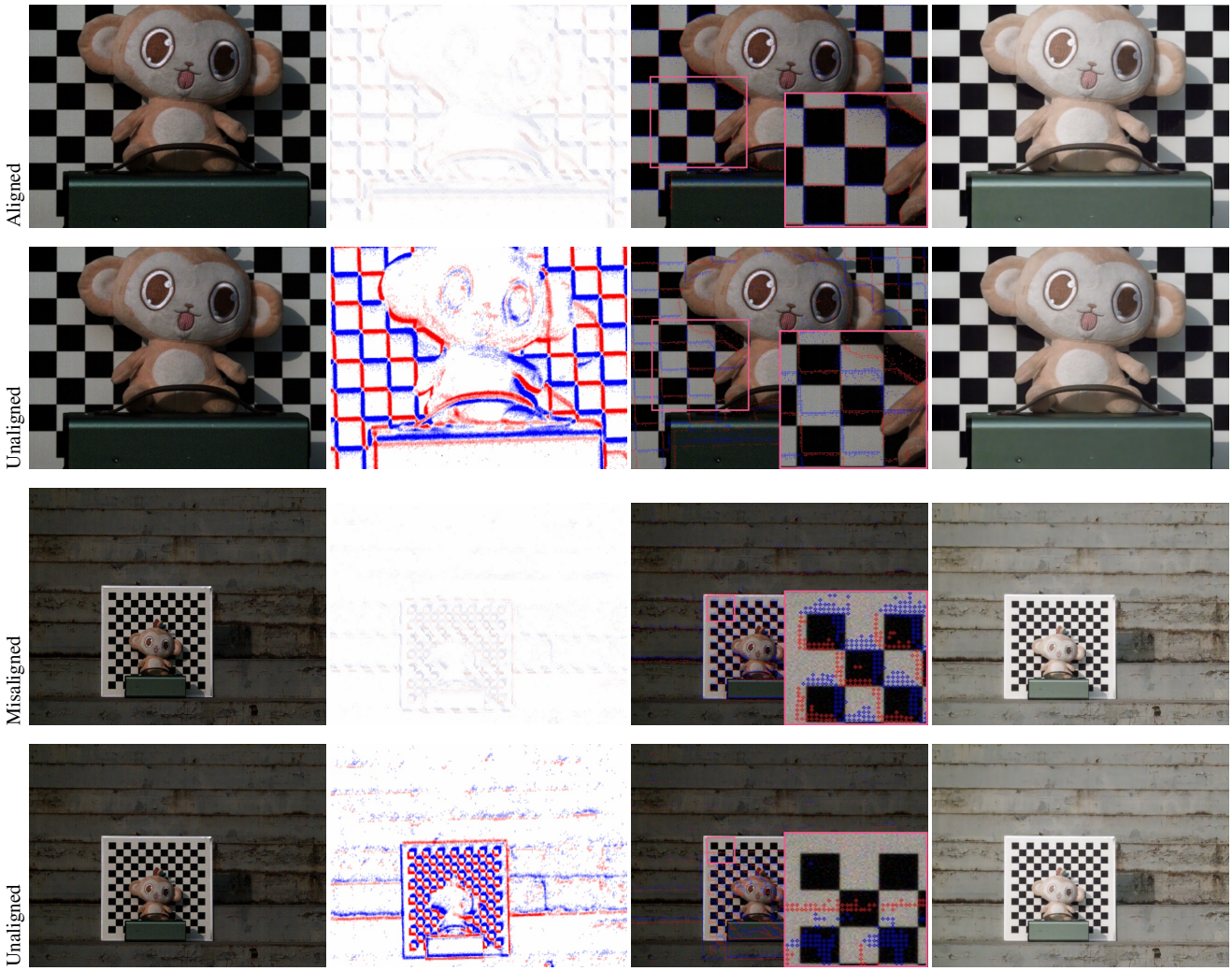
[18] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. URetinex-Net: Retinex-Based Deep Unfolding Network for Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2, 3, 7, 8, 9, 10, 11

[19] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-Aware Low-Light Image Enhancement. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 2, 3, 7, 8, 9, 10, 11

[20] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning Temporal Consistency for Low Light Video Enhancement From Single Images. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 2, 3, 7, 8, 9, 10, 11

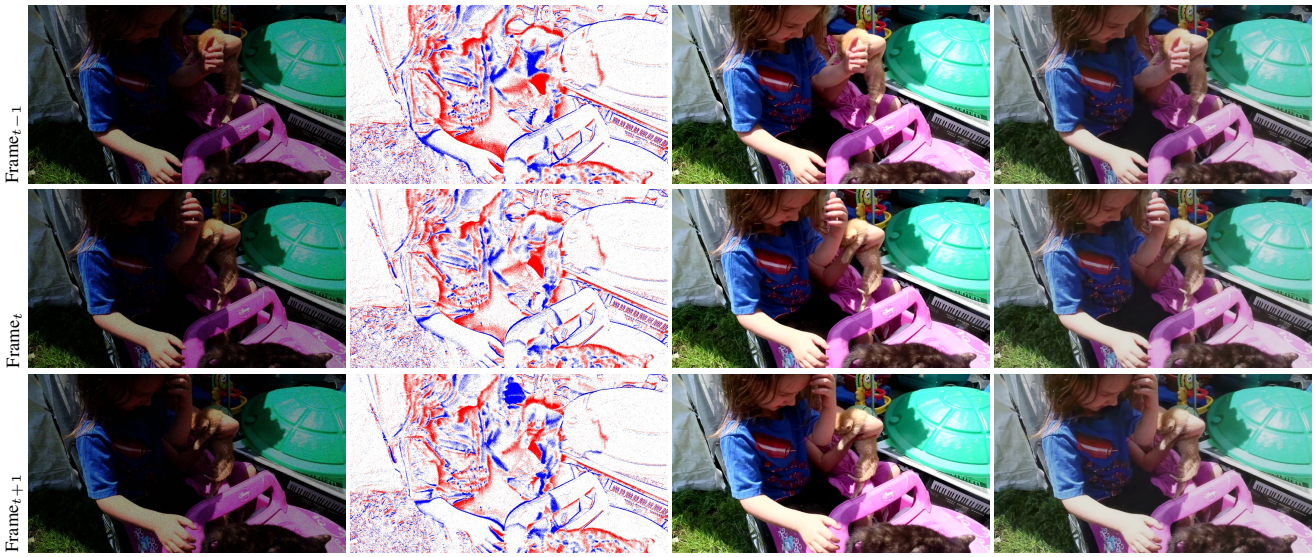
[21] Song Zhang, Yu Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to See in the Dark with Events. In *Proc. of European Conference on Computer Vision*, 2020. 1, 2, 3, 7, 8, 9, 10, 11

[22] Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, and Jiangbo Lu. Revisiting Temporal Alignment for Video Restoration. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 1



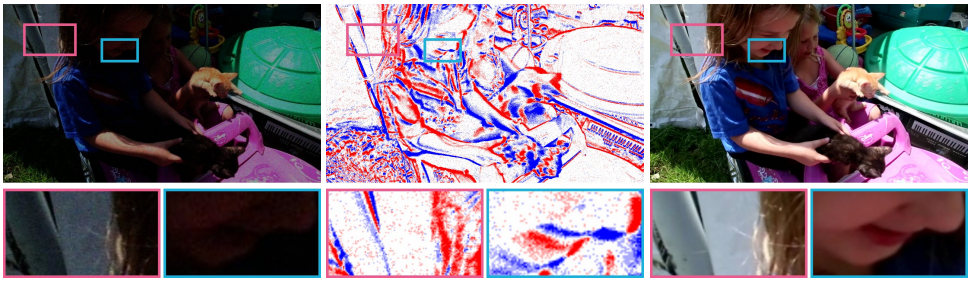
(a) Low-light frame (b) Events (c) Overlaid frames (a) and event (b) (d) Ours

Figure 7. Visual comparison results with different degree of misalignment.

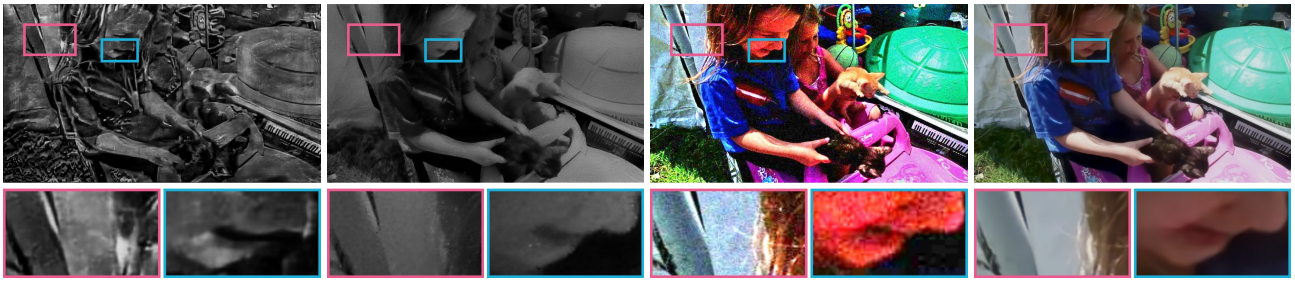


(a) Low-light frame (b) Events (c) Ground Truth (d) Ours

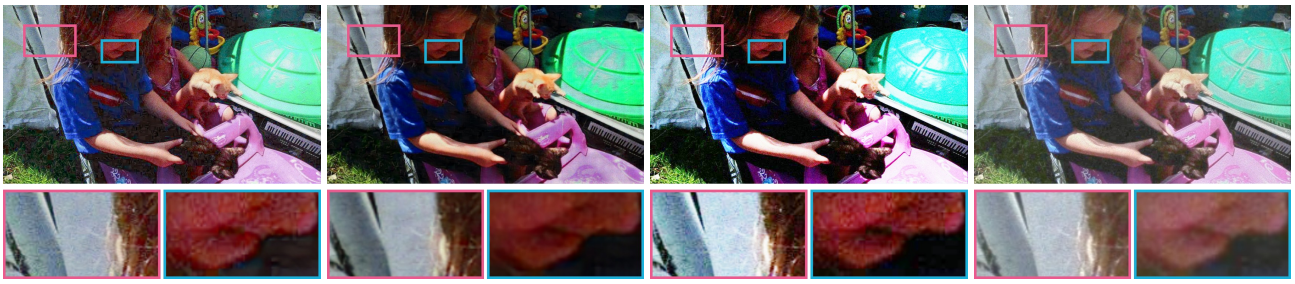
Figure 8. Visual demonstration of the robustness of the proposed method to discontinuity of motion ((a) *vs*. (b), the cat's head occluded) and lighting inconsistency ((a) *vs*. (c), the cat in shadow).



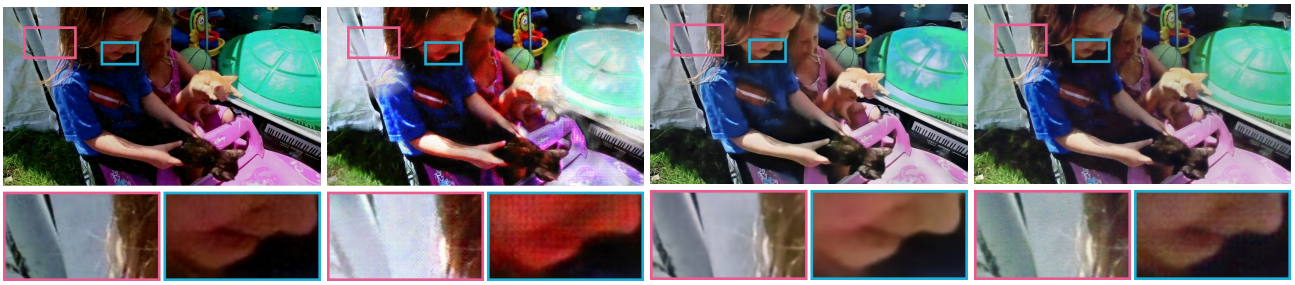
(a) Low-light frame (b) Events (c) Ground truth



(d) DVS-Dark [21] (e) E2VID [12] (f) LIME [5] (g) **Ours**

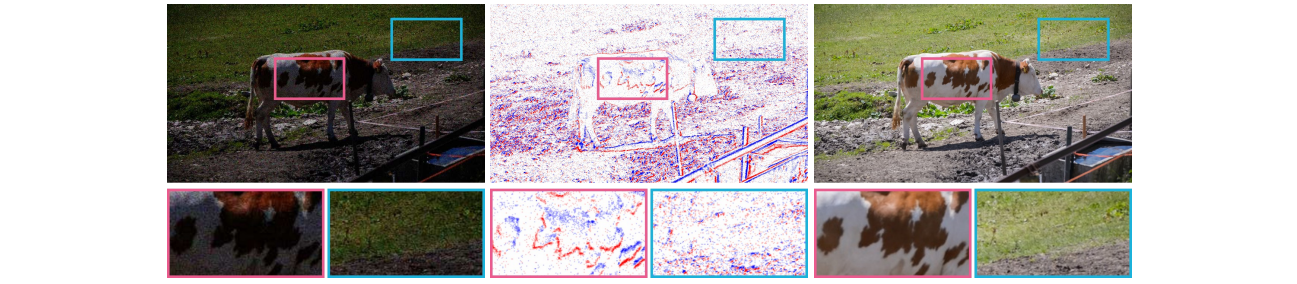


(h) URetinex-Net [18] (i) Transformer [19] (j) SCI [9] (k) StableLLVE [20]

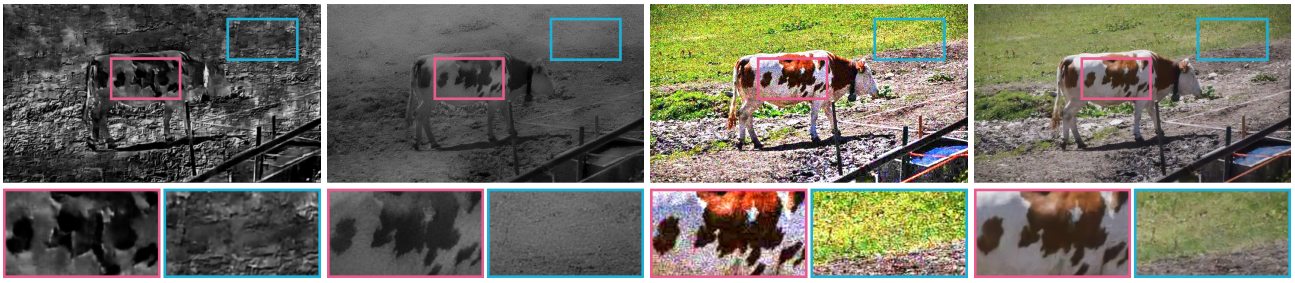


(l) MBLLEN [8] (m) SDSD [17] (n) MBLLEN† [8] (o) SDSD† [17]

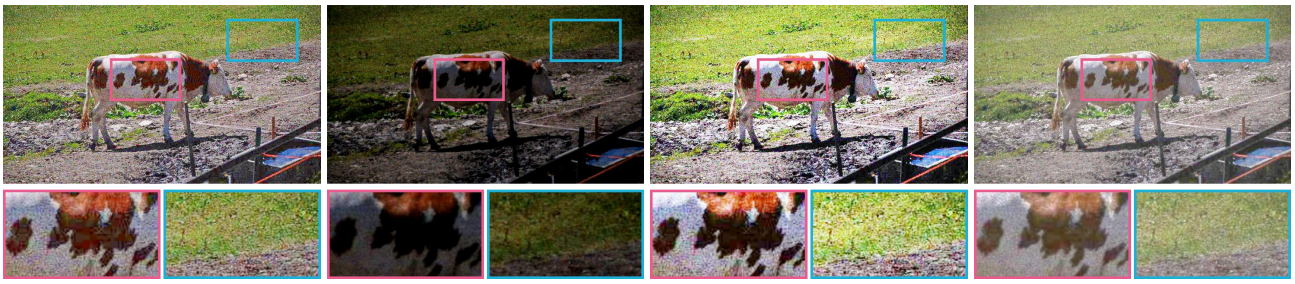
Figure 9. Visual comparison results with state-of-the-art methods on synthetic data.



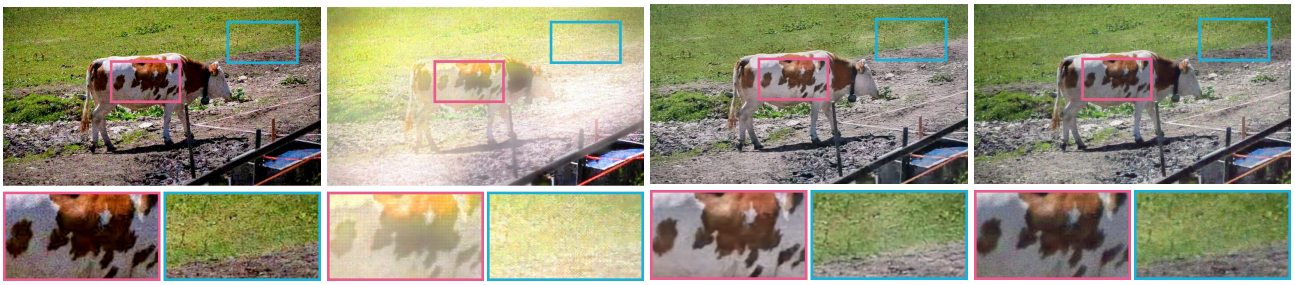
(a) Low-light frame (b) Events (c) Ground truth



(d) DVS-Dark [21] (e) E2VID [12] (f) LIME [5] (g) **Ours**

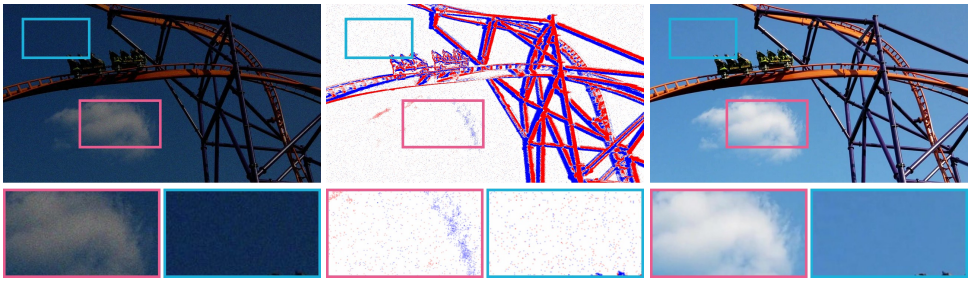


(h) URetinex-Net [18] (i) Transformer [19] (j) SCI [9] (k) StableLLVE [20]

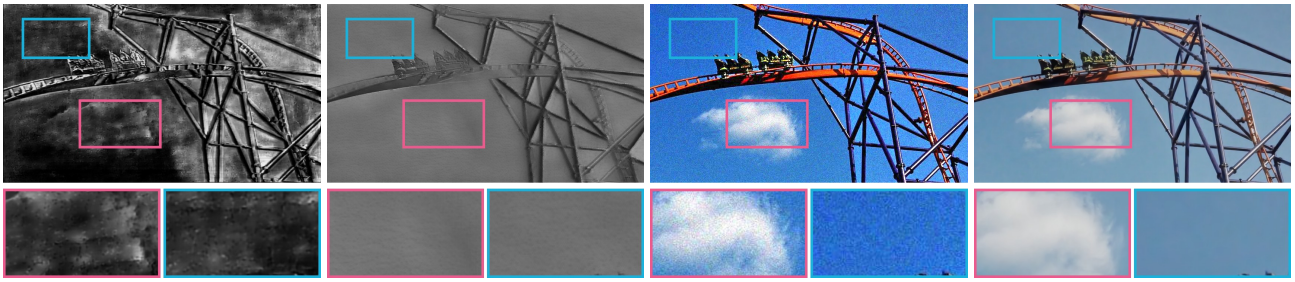


(l) MBLLEN [8] (m) SDSD [17] (n) MBLLEN† [8] (o) SDSD† [17]

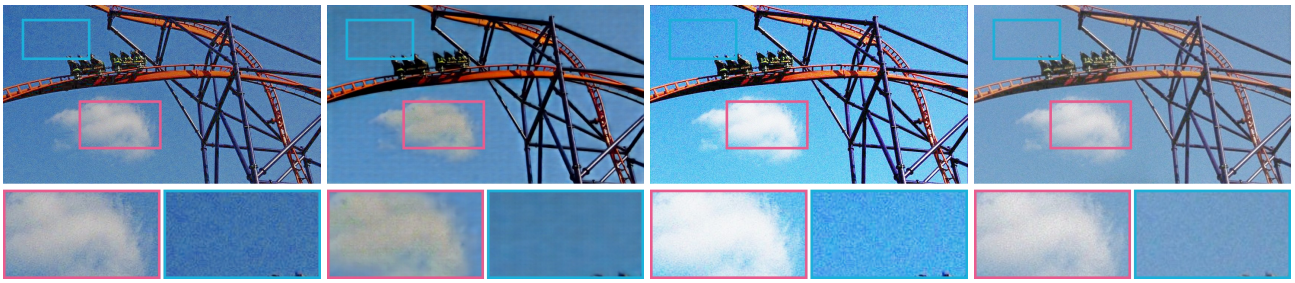
Figure 10. Visual comparison results with state-of-the-art methods on synthetic data.



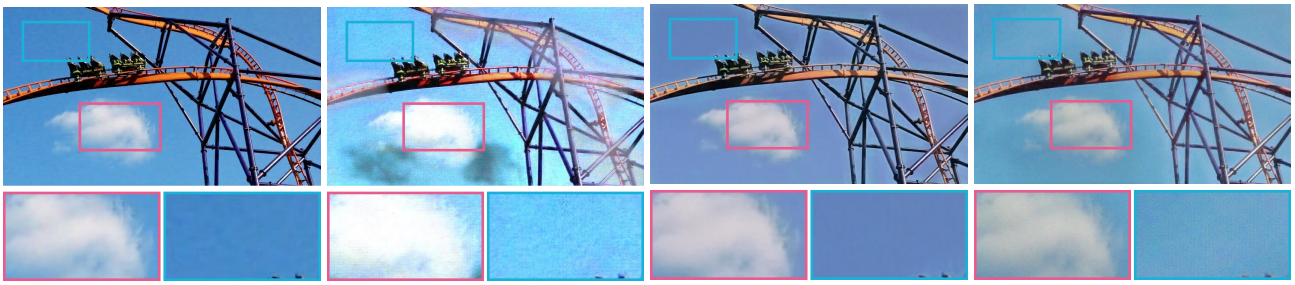
(a) Low-light frame (b) Events (c) Ground truth



(d) DVS-Dark [21] (e) E2VID [12] (f) LIME [5] (g) **Ours**

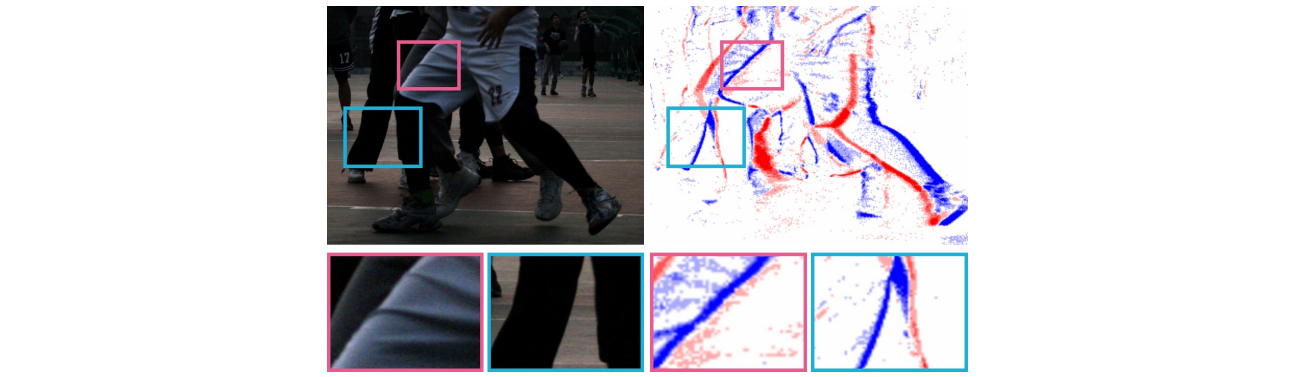


(h) URetinex-Net [18] (i) Transformer [19] (j) SCI [9] (k) StableLLVE [20]



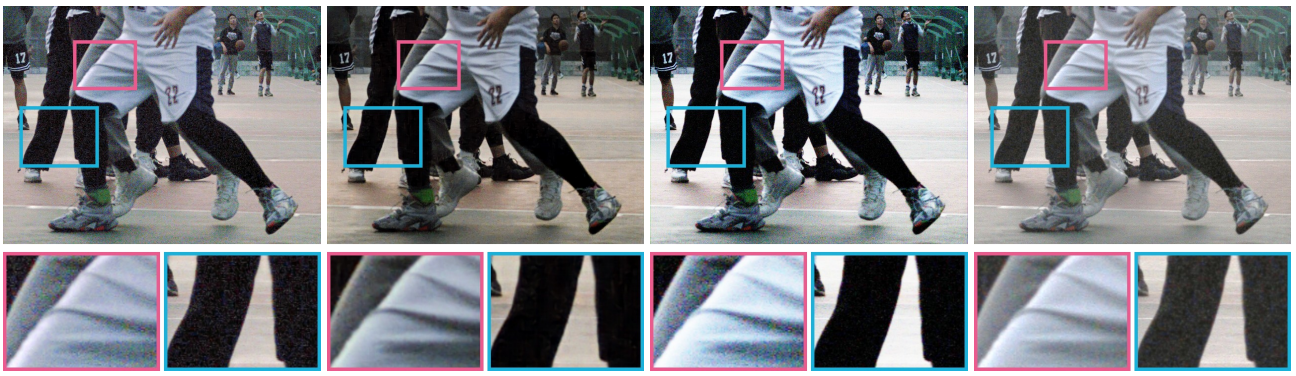
(l) MBLLEN [8] (m) SDSD [17] (n) MBLLEN† [8] (o) SDSD† [17]

Figure 11. Visual comparison results with state-of-the-art methods on synthetic data.

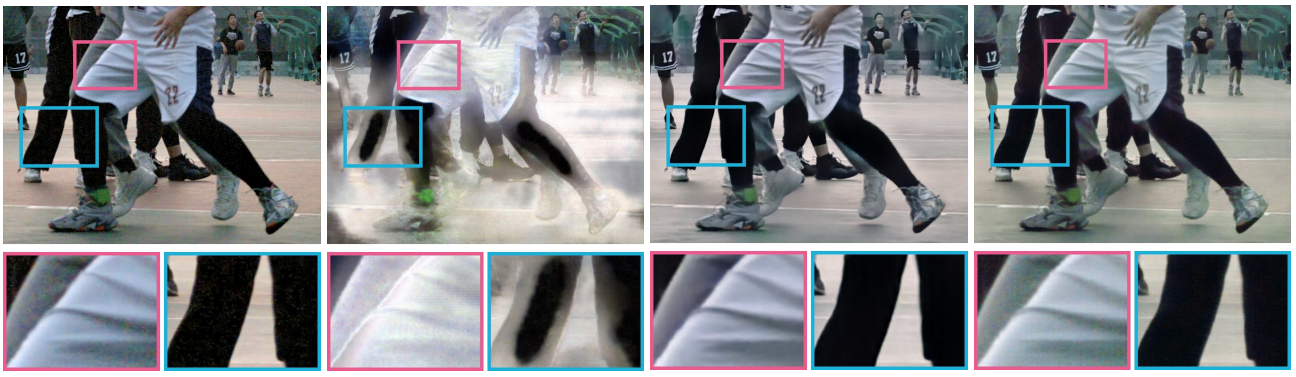


(a) Low-light frame (b) Events

(c) DVS-Dark [21] (d) E2VID [12] (e) LIME [5] (f) **Ours**

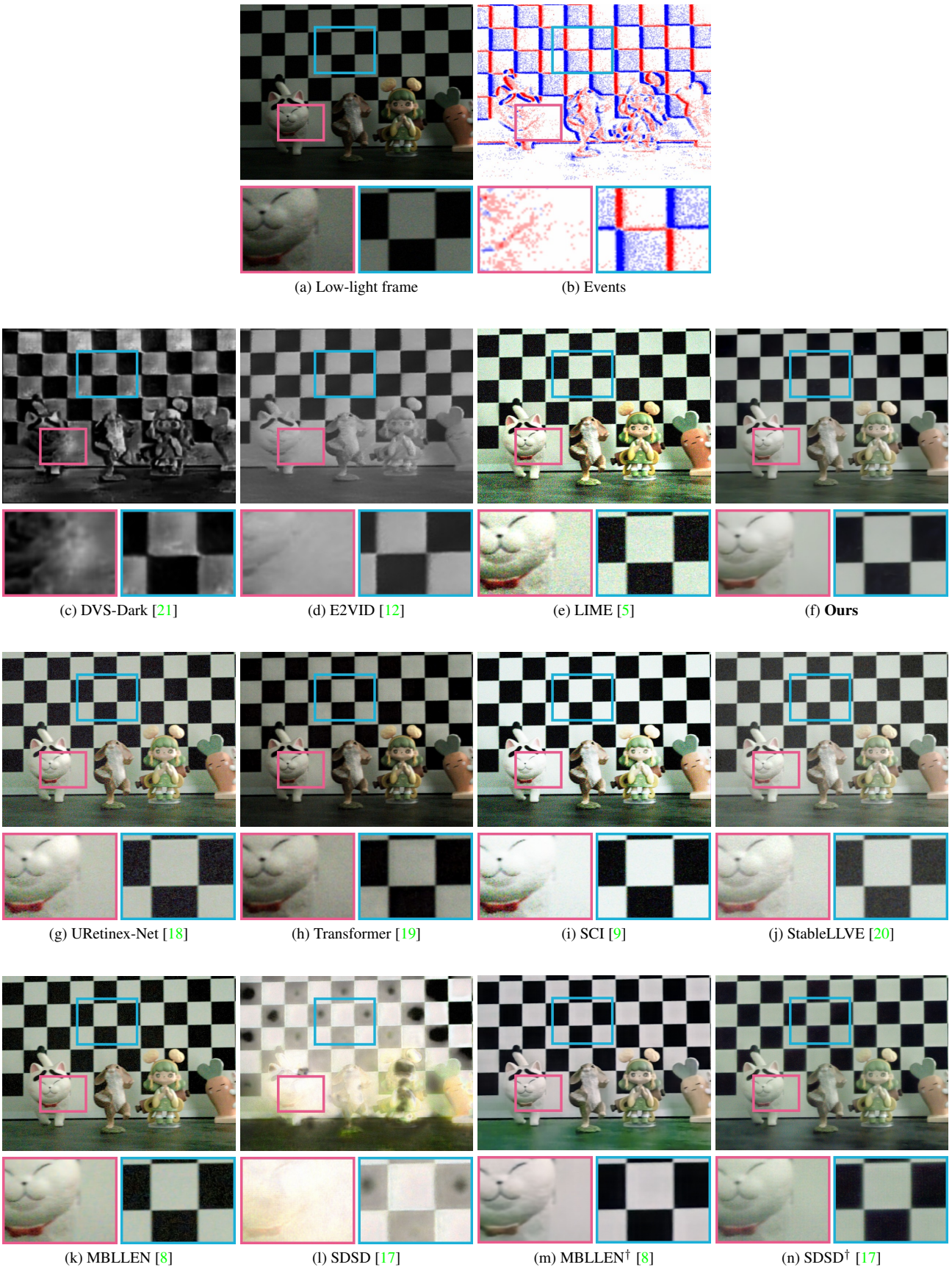


(g) URetinex-Net [18] (h) Transformer [19] (i) SCI [9] (j) StableLLVE [20]



(k) MBLLEN [8] (l) SDSD [17] (m) MBLLEN† [8] (n) SDSD† [17]

Figure 12. Visual comparison results with state-of-the-art methods on real data.



(a) Low-light frame (b) Events

(c) DVS-Dark [21] (d) E2VID [12] (e) LIME [5] (f) **Ours**

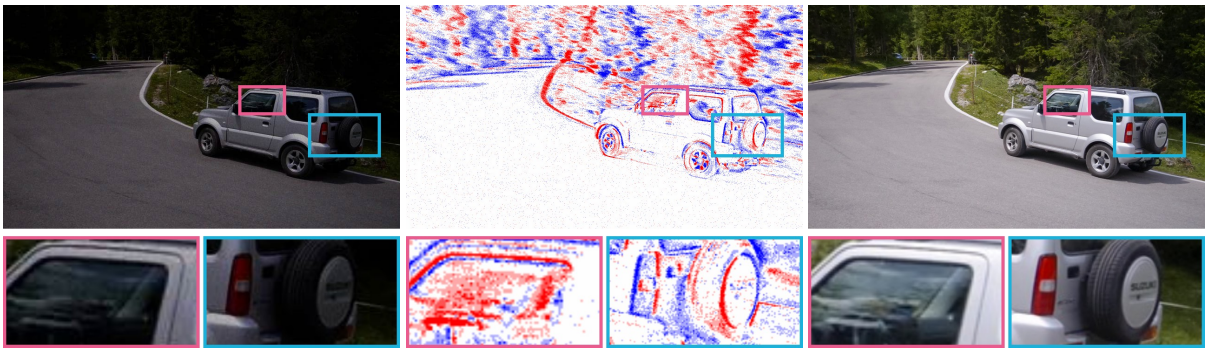
(g) URetinex-Net [18] (h) Transformer [19] (i) SCI [9] (j) StableLLVE [20]

(k) MBLLEN [8] (l) SDSD [17] (m) MBLLEN† [8] (n) SDSD† [17]

Figure 13. Visual comparison results with state-of-the-art methods on real data.

(a) Low-light frame (b) Events (c) Ours

Figure 14. Visual results on real data captured in DSEC [3] (the first four rows) and Time Lens [16] (the last two rows).

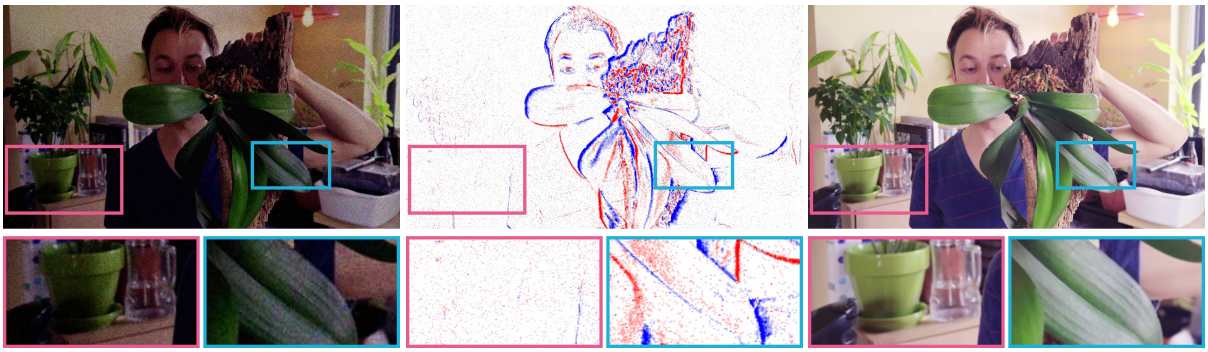


(a) Low-light frame (b) Events (c) Ground truth

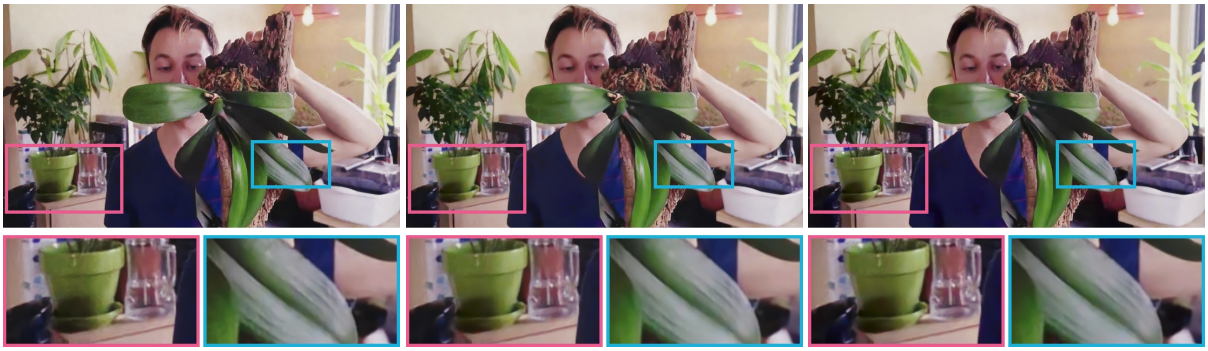
(d) w/o $\mathcal{F}_{\text{spat}}$ (e) w/o events (f) w/o feature alignment

(g) w/o $\mathcal{F}_{\text{temp}}$ (h) w/o noise simlulation (i) Ours

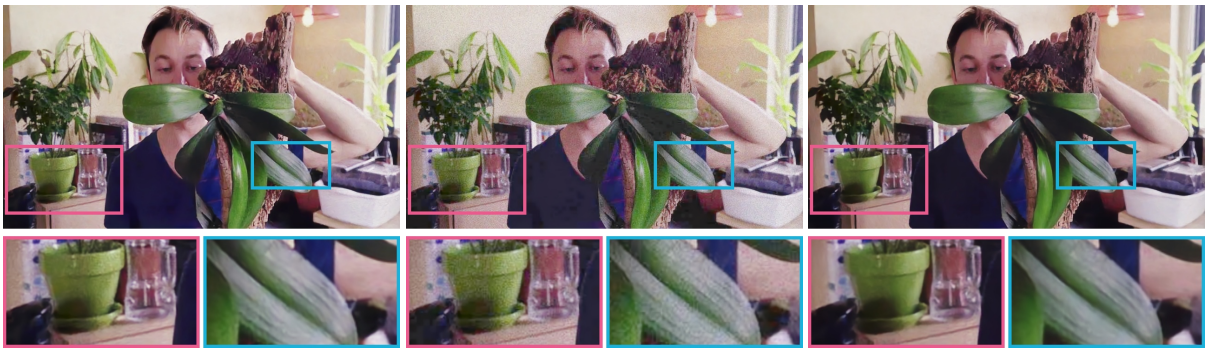
Figure 15. Qualitative comparison results on different cases of the ablation studies.



(a) Low-light frame (b) Events (c) Ground truth



(d) w/o $\mathcal{F}_{\text{spat}}$ (e) w/o events (f) w/o feature alignment



(g) w/o $\mathcal{F}_{\text{temp}}$ (h) w/o noise simlulation (i) Ours

Figure 16. Qualitative comparison results on different cases of the ablation studies.