Seeing Motion at Nighttime with an Event Camera

Supplementary Material

Summary

The supplementary material is organized as follows.

- Section 1 introduces the implementation of NER-Net and experimental settings.
- Section 2 discusses more ablation studies of the proposed method.
- Section 3 shows more visualization results on real nighttime dynamic scene datasets.

1. Implementation Details

1.1. Non-uniform Illumination Aware Module

We design Non-uniform Illumination Aware Module (NIAM) encoders comprising Global Context Block (GCB), Local Adaptation Gate (LAG), and Spatiotemporal Aggregation Unit (SAU) modules. The equations of NIAM are shown as follows

$$
\begin{aligned}
&\mathcal{X}_t = GCB(x_t)\\
&g_t = \tanh(\mathcal{W}_{xg} * \mathcal{X}_t + \mathcal{W}_{hg} * \mathcal{H}^l_{t-1} + b_g)\\
&i_t = \sigma(\mathcal{W}_{xi} * \mathcal{X}_t + \mathcal{W}_{hi} * \mathcal{H}^l_{t-1} + b_i)\\
&f_t = \sigma(\mathcal{W}_{xf} * \mathcal{X}_t + \mathcal{W}_{hf} * \mathcal{H}^l_{t-1} + b_f)\\
&f^{LAG}_t = \sigma(f_t - \alpha i_t)\\
&\mathcal{C}^l_t = f^{LAG}_t \odot \mathcal{C}^{l-1}_t + i_t \odot g_t\\
&g_t{}' = \tanh(\mathcal{W}'_{xg} * \mathcal{X}_t + \mathcal{W}_{mg} * \mathcal{M}^{l-1}_t + b'_g)\\
&i_t{}' = \sigma(\mathcal{W}'_{xi} * \mathcal{X}_t + \mathcal{W}_{mi} * \mathcal{M}^{l-1}_t + b'_i)\\
&f_t{}' = \sigma(\mathcal{W}'_{xf} * \mathcal{X}_t + \mathcal{W}_{mf} * \mathcal{M}^{l-1}_t + b'_f)\\
&\mathcal{M}^l_t = f'_t \odot \mathcal{M}^{l-1}_t + i'_t \odot g'_t\\
&o_t = \sigma(\mathcal{W}_{xo} * \mathcal{X}_t + \mathcal{W}_{ho} * \mathcal{H}^l_{t-1} + \mathcal{W}_{co} * \mathcal{C}^l_t + \mathcal{W}_{mo} * \mathcal{M}^l_t + b_o)\\
&\mathcal{H}^l_t = o_t \odot \tanh(\mathcal{W}_{1\times1} * [\mathcal{C}^l_t, \mathcal{M}^l_t]),
\end{aligned}
\tag{1}
$$

where $\mathcal{C}^l_t$ is the temporal cell including GAB and LAG, which is updated repeatedly over time. $\mathcal{M}^l_t$ is the spatiotemporal memory that can aggregate hierarchical features across layers, and it also transmits the spatiotemporal memory from the top layer at time $t-1$ with rich semantic information to the bottom layer at time t through a progressive upsampling. In addition to the temporal memory $\mathcal{C}^l_t$, the spatiotemporal memory unit $\mathcal{M}^l_t$ can integrate and transmit information across layers.

1.2. Event Trail Suppression

The details of the proposed event trail suppression (ETS) method is in Algorithm 1. The input of ETS is a set of events (x, y, t, p), and the output of ETS is a set of calibration events (x, y, t', p). Note that, ETS solely corrects the timestamps of events. T_i is the timestamps set of input events, P_i is the polarities set of input events, t^{first}_i is the timestamp of the first event. We empirically set $thr = 1s$, and $t_{interval} = 1\mu s$. The ablation study of ETS hyper-parameters is shown in Table 1.

Algorithm 1 Event Trail Suppression

for each pixel (x_i, y_i) **do**
  $T_i = [t^0_i, t^1_i, ..., t^n_i]$
  $P_i = [p^0_i, p^1_i, ..., p^n_i]$
  for $j = 1, 2, ..., n$ **do**
    Condition A: $p^n_i = p^{n-1}_i$
    Condition B: $t^n_i - t^{n-1}_i > t^{n-1}_i - t^{n-2}_i$
    Condition C: $t^n_i - t^{n-1}_i < thr$
    if A is True and B is True and C is True **then**
      $t^n_i = t^{first}_i + t_{interval}$
    end if
    update event$x^n_i, y^n_i, t^{'n}_i, p^n_i$
  end for
end for

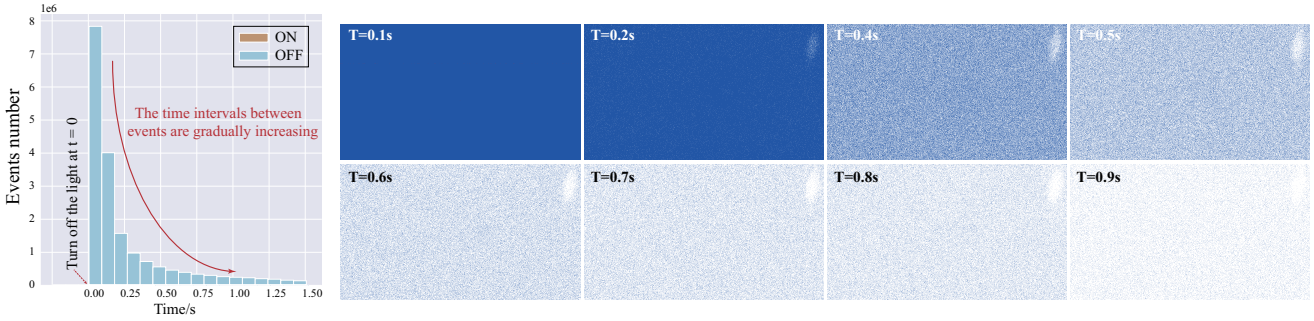
1.3. Datasets Details

• **RLED.** We uniformly selected 100 image sequences from the RLED dataset for training and 23 image sequences for testing, with each sequence containing 99 images. The optical flow between two frames is estimated using Flow-former [4] and utilized to compute the temporal consistency loss.

• **DSEC-night.** DSEC [2] is a large-scale real-world dataset, including adverse conditions such as fast-moving objects and precarious illumination from daytime to nighttime. The event sensor is Prophesee Gen3.1 (640*480). We choose 4 nighttime sequences of DESC (zurich_city_09_b, zurich_city_09_c, zurich_city_09_d, zurich_city_09_e) to compose a test set namely DSEC-night, containing 3762 images. And then we align events and images based on the camera calibration parameters provided by the authors.

• **MVSEC-night.** MVSEC [12] is a real-world dataset for 3D perception, and the paired data of events and images was provided by a DAVIS346 (346*260) camera. We extract the nighttime driving data (outdoor_night1) for testing, comprising 2472 images.

• **VECtor-hdr.** VECtor [1] is a versatile event-centric benchmark for multi-sensor Simultaneous Localization And Mapping (SLAM). The event stereo cameras have VGA resolution (Prophesee Gen3, 640*480) with a horizontal baseline of about 17 cm. We selected two scenes with challenging lighting conditions (hdr-normal and hdr-fast) for testing.



(a) Temporal distritbution of events

(b) Visualizing events across different time intervals

Figure 1. Trailing events statistical characteristics. (a) represents the statistical histogram of events and (b) represents the visualization of events across different time intervals. It can be observed that the change rate of OFF events gradually decreases after the lights are turned off.

2. Ablation Study and Discussion

2.1. Trailing Events Statistical Characteristics

To verify the temporal distribution characteristics of trailing events, we design a flashlight experiment. We place a lamp and an event camera inside a dark chamber, with the lamp continuously on. Subsequently, we turn off the lamp and record the distribution of events. The statistical characteristic of events is shown in Fig. 1. The time intervals between events are gradually increasing after an excitation, this phenomenon aligns with our assumption regarding the characteristics of trailing events and motivates us to design the ETS algorithm.

2.2. Sensitivity of ETS hyper-parameters

We study the sensitivity of ETS hyper-parameters in Table. 1. First, we fix $t_{interval} = 1\mu s$, the thr are set at 0.1s, 0.5s, and 1s, respectively. The results are approximately equal when $thr = 1$s and $thr = 0.5$s, whereas there is a significant decrease when $thr = 0.1$s. The reason is that the length of trailing events is typically greater than 0.1s, hence, excessively low thresholds prematurely terminate the trail suppression process. Then, we fix $thr = 1$s, the $t_{interval}$ are set at 1μs, 5μs and 10μs. It's noticeable that as $t_{interval}$ increases, there's a slight decline in the quality of reconstruction. Thus, we set $thr = 1$s and $t_{interval} = 1\mu$s.

2.3. Effect of the Temporal Loss

Table 2 compares the results between w/ and w/o temporal consistency loss. By employing TC loss, the NER-Net can better preserve the continuity of information, consequently enhancing the reconstruction performance.

2.4. Generalizability during Daytime

We capture 22 sequences using the same system without an ND filter namely real daytime event dataset (RDED), and created training sets in three ways: (1) normal-light data only; (2) low-light data only; (3) both datasets mixed in a 1:1 ratio. As shown in Table 3, both training and testing on the same data have better results. The divergence in data distribution between training and testing sets may lead to a reduction in reconstruction quality. Combining training on two types of data can alleviate this situation.

Fig. 2 illustrates the generalization in daytime scenarios of the proposed method. Table 4 reports the quantitative results. Besides, the NER-Net trained by RDED can generate natural results on other unseen daytime datasets, including DSEC [2], MVSEC [12], IJRR [6], and HQF [7]. This indicates that training the model with real-world data can lead to better generalization. The suboptimal performance of NER-Net on the IJRR data is attributed to the significant resolution gap between its training set and the IJRR (1280*720 vs 240*180).

Paramters		RLED		
		MSE ↓	SSIM ↑	LPIPS ↓
thr	0.1s	0.018	0.710	0.364
	0.5s	0.012	0.715	0.309
	1s	**0.011**	**0.717**	**0.309**
$t_{interval}$	1μs	**0.011**	**0.717**	**0.309**
	5μs	0.012	0.717	0.311
	10μs	0.015	0.714	0.313

Table 1. Ablation studies of ETS hyper-paramters.

	MSE ↓	SSIM ↑	LPIPS ↓
w/o temporal loss	0.012	0.713	0.312
w/ temporal loss	**0.011**	**0.717**	**0.309**

Table 2. Effect of the Temporal Loss.

2.5. The Improvement to Hybrid Methods

The proposed method also can enhance the effectiveness of hybrid methods in nighttime imaging. NeurImg-HDR

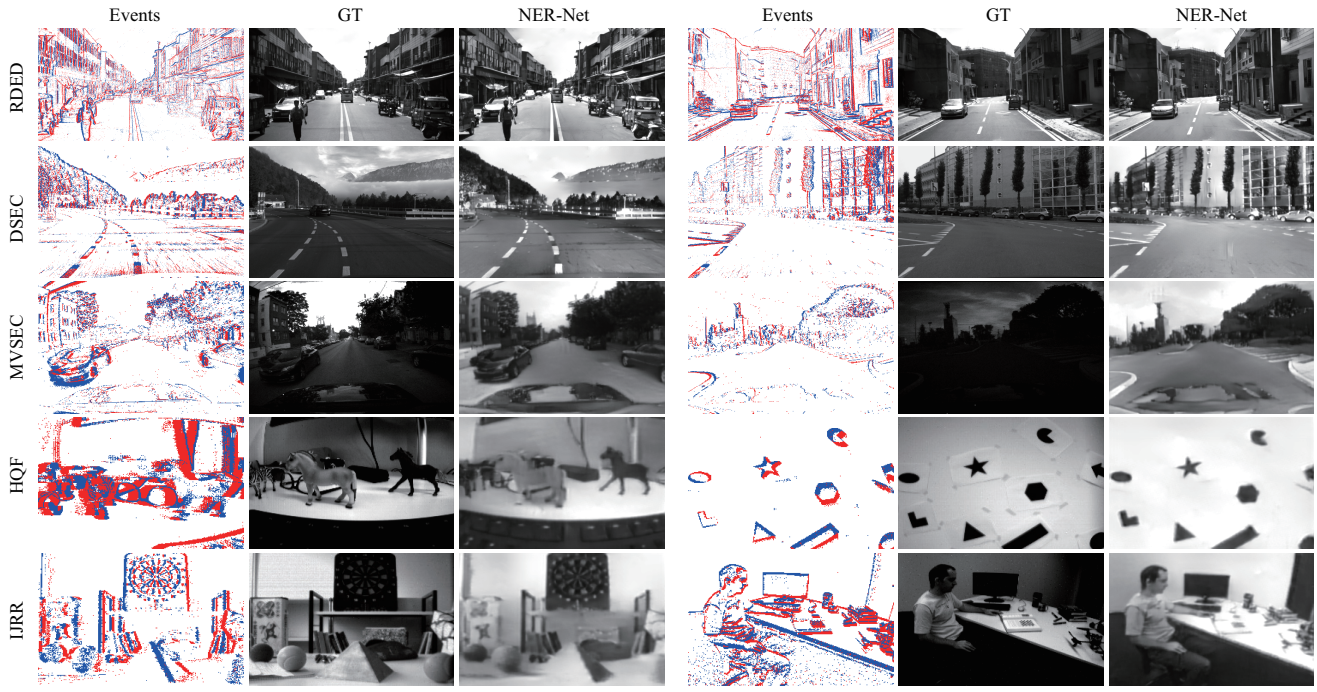


Figure 2. Generalization in real daytime datasets. The proposed method demonstrates strong generalization capabilities in daytime scenarios, delivering excellent reconstruction performance on unseen datasets.

Training data	Normal-light			Low-light		
	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓
Normal	0.008	0.724	0.301	0.065	0.387	0.541
Low	0.071	0.427	0.526	0.012	0.711	0.315
Normal+Low	0.010	0.713	0.312	0.014	0.703	0.323

Table 3. Quantitative results on the daytime and nighttime data.

Methods	IJRR			DSEC			RDED		
	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓
E2VID+	0.065	0.566	**0.343**	0.080	0.308	0.606	0.082	0.435	0.518
ET-Net	**0.050**	**0.592**	0.345	0.084	0.266	0.625	0.081	0.426	0.545
DVS-Dark	0.087	0.354	0.424	0.108	0.184	0.694	0.071	0.372	0.562
NER-Net(ours)	0.068	0.589	0.348	**0.070**	**0.337**	**0.589**	**0.018**	**0.727**	**0.309**

Table 4. Quantitative results on daytime datasets.

[3] is a SOTA HDR imaging method that fuses events and images. We replace the intensity map reconstruction module in NeurImg-HDR from pre-trained E2VID to pre-trained NER-Net and test it on the DSEC-night and MVSEC-night datasets without any fine-tuning. The quantitative and qualitative results are illustrated in Table 5 and Fig. 4. It can be observed that NER-Net effectively reduces reconstruction artifacts and improves image quality.

2.6. Limitation

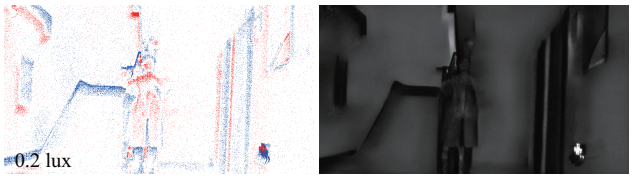
Imaging in extremely dark motion scenes (*e.g.* <0.5 lux) remains challenging. On one hand, the brightness variations within the scene may fall below the event camera's triggering threshold, leading to the loss of fine texture details. On the other hand, the event camera experiences a sharp decline in signal-to-noise ratio, making it difficult to reconstruct reasonable scene intensities, as shown in Fig. 3.

Method	LOE ↓	NIQE ↓	SPAQ ↑
with E2VID	126.02	17.12	7310.06
with NER-Net	**125.82**	**17.07**	**11213.68**

Table 5. The improvement to hybrid methods.

Input size	Parameters (M)	Memory (MB)	Flops (GFlops)	Time(ms)
346 * 260	19.36	2066	83.16	10.35
560 * 400	19.36	2492	200.52	13.50
1120 * 660	19.36	4152	672.24	41.82

Table 6. Computational performance.



(a) Event frames (b) Reconstruction results

Figure 3. Limitation of the proposed method. The event camera failed to capture scene texture details under extremely low-light conditions (about 0.2 lux), resulting in distortion in the reconstructed scene intensity.

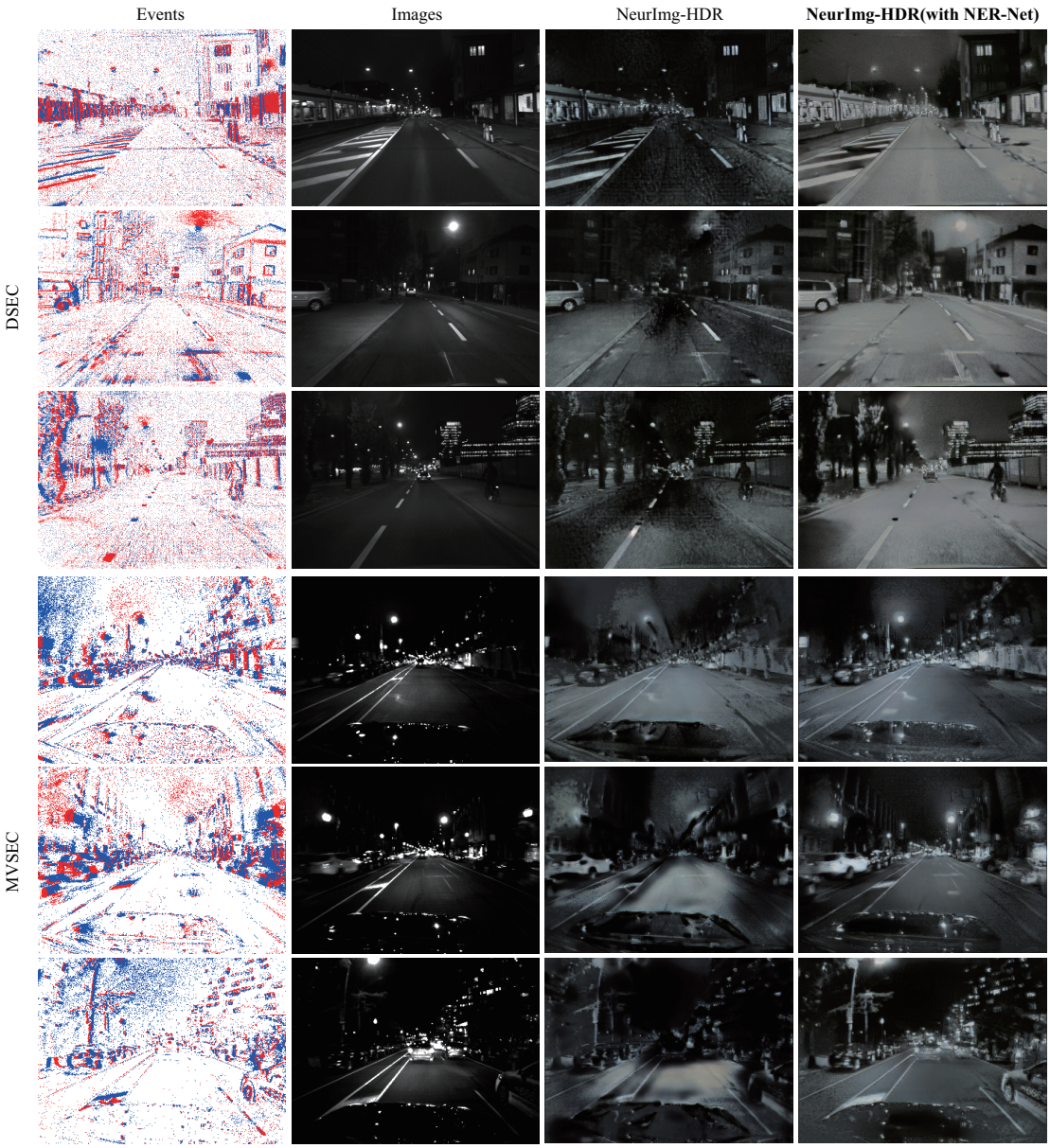


Figure 4. NeurImg-HDR reconstruction results. Replacing E2VID with NER-Net significantly improves the quality of images.

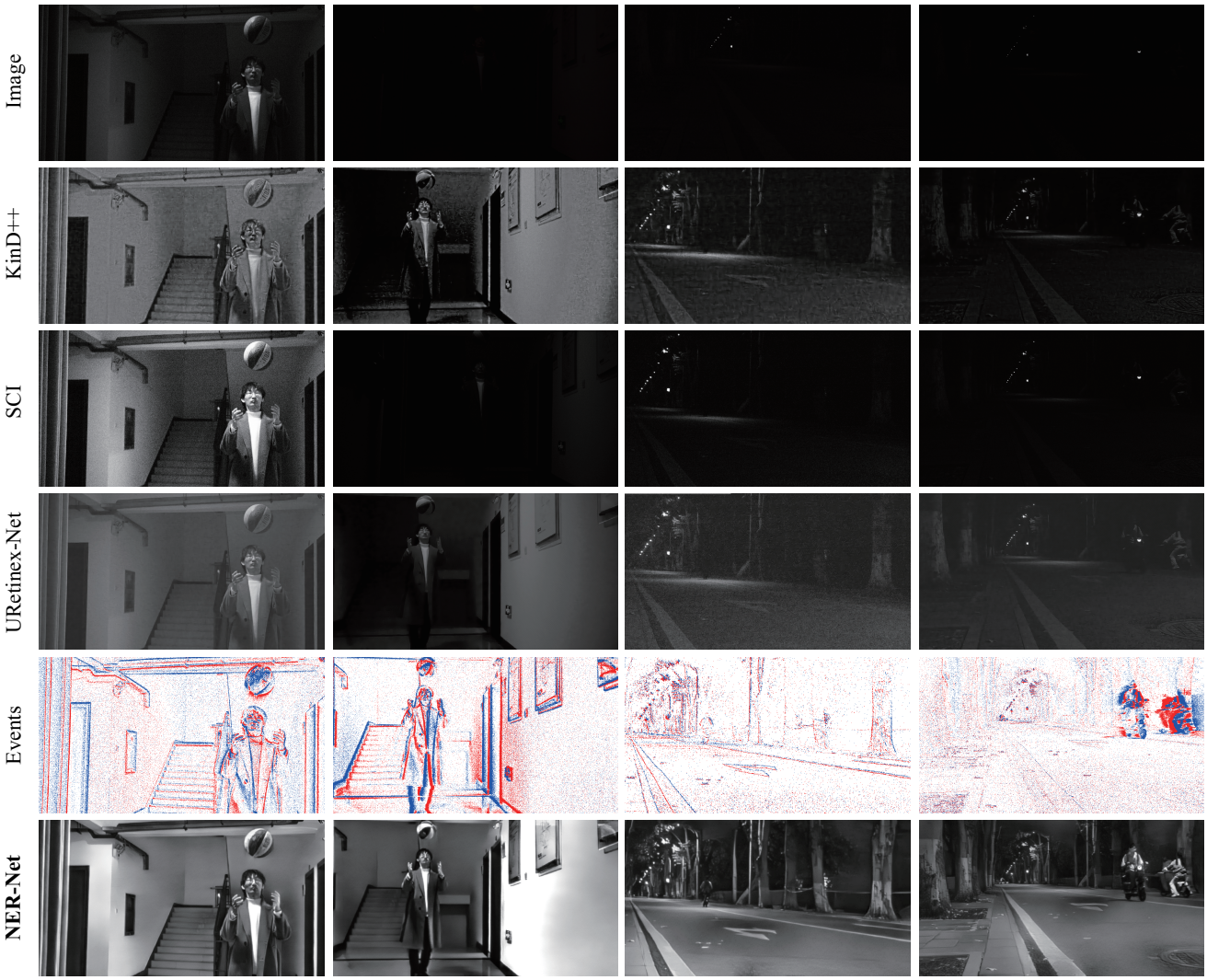
2.7. Computational Performance

We using an NVIDIA A100 GPU for all experiments. Table 6 shows the details of computational performance on NER-Net on RLED (1120 * 660), DSEC (560 * 400), and MVSEC (346 * 260) datasets respectively.

3. Additional Results

3.1. Comparison with image enhancement methods

We compared the proposed NER-Net with three state-of-the-art low-light enhancement methods (KinD++ [11], SCI [5], and URetinex-Net [9]) across various light levels and different motion speeds. We carefully set exposure times based on object motion speeds to prevent motion blur in images. The lens aperture for both the event camera and frame camera is set to F2.0. To maintain visual consistency, we convert the enhanced results into grayscale images. Note that, due to the non-uniformity of artificial light at night, we employ an illuminance meter to measure the average illuminance in the imaging system. The illumination values may vary across different regions within the field of view. When capturing moderately paced moving objects at 20.0 lux (such as throwing a ball), we set the exposure time to 10ms, and the low-light enhancement methods work effectively, as



(a) Throwing a ball (b) Throwing a ball (c) Riding (d) Fast riding

Figure 5. Qualitative results on challenging nighttime dynamic scenes. (a) Throwing a ball, exposure time = 10 ms, scene illumination = 20.0 lux, low ISO. (b) Throwing a ball, exposure time = 10 ms, scene illumination = 10.4 lux, low ISO. (c) Riding, exposure time = 10 ms, scene illumination = 6.7 lux, high ISO. (d) Fast riding, exposure time = 5 ms, scene illumination = 6.7 lux, low ISO.

shown in Fig. 5(a). In Fig. 5(b), the staircase on the left side of the image is further away from the light source and the enhanced image still fails to recover the scene information. In Fig. 5(c), we refrain from adjusting the exposure time and capture images from a greater distance from the imaging device to prevent motion blur. We increase the ISO of the camera to accommodate darker scenes (6.7 lux), however, the enhanced image appears noisy. Fig. 5(d) depicts a more challenging scenario with fast-moving objects, hence, we reduce the exposure time to 5ms. When ambient light diminishes and object motion speeds increase, the efficacy of low-light enhancement significantly diminishes. The excessively short exposure time hinders the frame-based camera from capturing adequate scene information, making recovery challenging even with enhancement methods. The proposed NER-Net sustains better imaging quality across varying light levels and different motion speeds and maintains good robustness across the aforementioned scenarios.

3.2. Comparison with event reconstruction methods

Additional qualitative results on RLED and DSEC [2] datasets are shown in Fig. 6. E2VID+ [7] and ET-Net [8] suffer from severe artifacts due to the substantial gap between the simulated and real-world data distributions. DVS-Dark [10] also fails to reconstruct natural images in nighttime dynamic scenes. NER-Net outperforms state-of-the-art methods in terms of visual quality and generalization ability on real-world nighttime datasets.

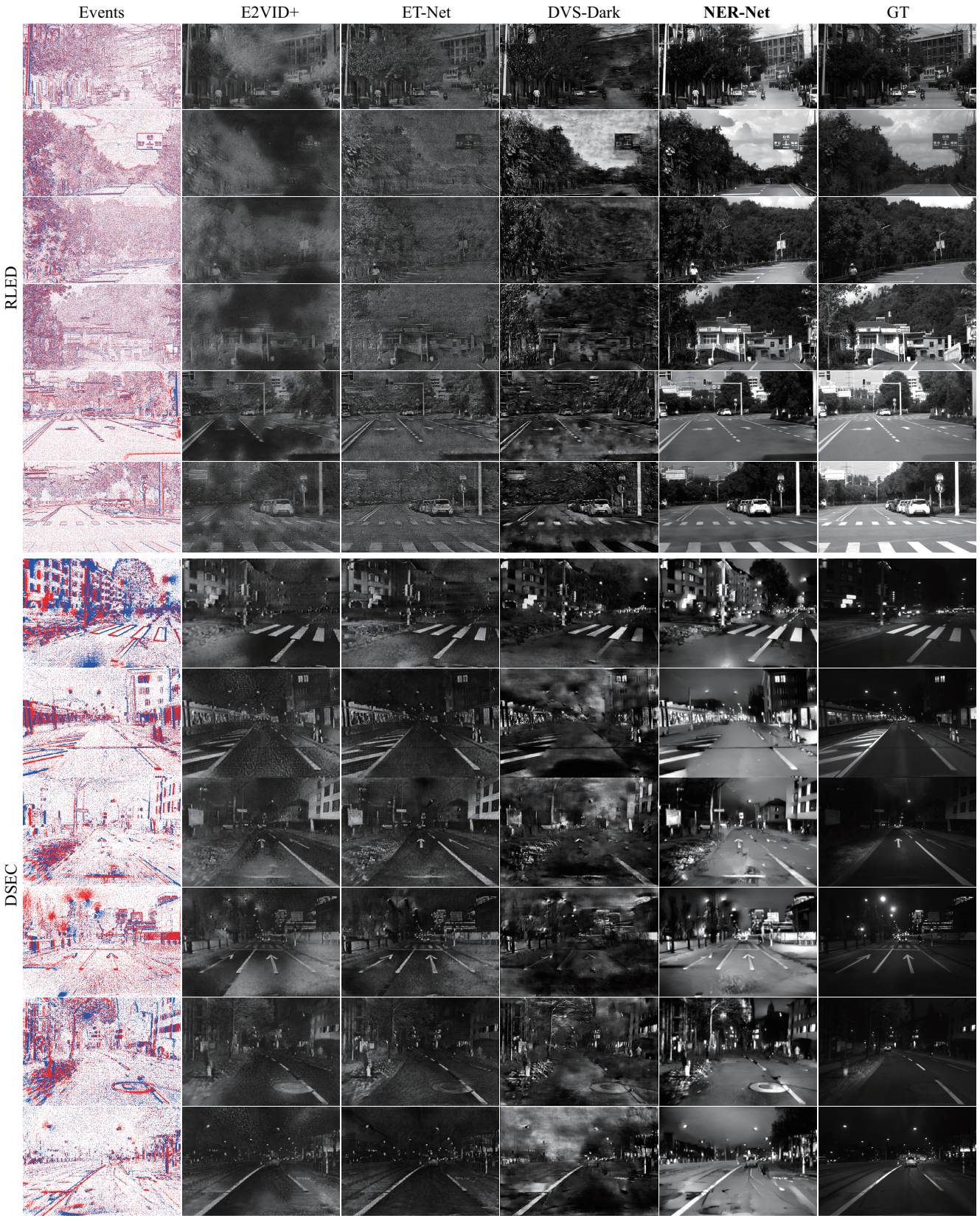


Figure 6. Qualitative results on real-world datasets. Other state-of-the-art methods fail to adapt to the distribution of nighttime events, resulting in severe reconstruction artifacts. NER-Net can reconstruct natural HDR images and still demonstrates strong generalization on unseen real-world nighttime datasets.

References

[1] Ling Gao, Yuxuan Liang, Jiaqi Yang, Shaoxun Wu, Chenyu Wang, Jiaben Chen, and Laurent Kneip. Vector: A versatile event-centric benchmark for multi-sensor slam. *IEEE Robot. Autom.*, 7(3):8217–8224, 2022. 1

[2] M.s Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robot. Autom.*, 6(3):4947–4954, 2021. 1, 2, 5

[3] J Han, Y Yang, P Duan, C Zhou, L Ma, C Xu, T Huang, I Sato, and B Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 3

[4] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. Cheung, H. Qin, J. Dai, and H. Li. Flowformer: A transformer architecture for optical flow. In *Eur. Conf. Comput. Vis.*, pages 668–685, 2022. 1

[5] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo. Toward fast, flexible, and robust low-light image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5637–5646, 2022. 4

[6] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *Int. J. Rob. Res.*, 36(2):142–149, 2017. 2

[7] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, Li. Kleeman, and R. Mahony. Reducing the sim-to-real gap for event cameras. In *Eur. Conf. Comput. Vis.*, pages 534–549, 2020. 2, 5

[8] W Weng, Y. Zhang, and Z. Xiong. Event-based video reconstruction using transformer. In *Int. Conf. Comput. Vis.*, pages 2563–2572, 2021. 5

[9] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5901–5910, 2022. 4

[10] S. Zhang, Y. Zhang, Z. Jiang, D. Zou, J. Ren, and B. Zhou. Learning to see in the dark with events. In *Eur. Conf. Comput. Vis.*, pages 666–682, 2020. 5

[11] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang. Beyond brightening low-light images. *Int. J. Comput. Vis.*, 129:1013–1037, 2021. 4

[12] A. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robot. Autom.*, 3(3):2032–2039, 2018. 1, 2