Supplementary Material for "LED: A Large-scale Real-world Paired Dataset for Event Camera Denoising"

Yuxing Duan[1]
[1]National Key Lab of Multispectral Information Intelligent Processing Technology
[1]Huazhong University of Science and Technology

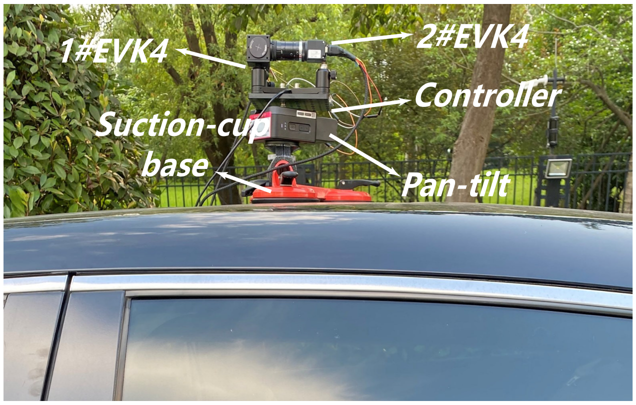

Figure 1. Outdoor vehicle-mounted data collection Platform.

Summary
This supplementary material is organized as follows:

- Sec. 1 discusses more analysis about DED Framework in the main paper Sec.3;
- Sec. 2 introduces the implementation details of the proposed network in the main paper Sec.4;
- Sec. 3 and Sec. 4 shows more comparative results of different datasets and methods.

1. DED Framework

1.1. Spatiotemporal Calibration

Here, we describe the spatiotemporal calibration procedure between two 1280*720 resolution event cameras of Prophesee EVK4 with a 16 mm lens. The two cameras share a common field of view through a beam splitter (Thorlabs CM1- DCH/M) with a 50% splitting. Although they are physically located in the same position, stereo geometric calibration is still needed to resist some inevitable factors such as slight mechanical deformation. During the calibration stage, both cameras remain stationary. To generate an event stream, a checkerboard pattern is manually moved within the camera field of view. To form event images, the event stream within a short time window is reconstructed into intensity images with E2VID [9]. The key points extracted from the corners of the checkerboard image are used to estimate the rotation and translation matrices. After stereo geometric calibration, the spatial registration of the two cameras on the 2D plane is more accurate.

The temporal registration procedure is developed based on STM32, which can achieve a temporal resolution of μs level and thus realize synchronous recording requirement within the cumulative time window of ms.The calibrated device for outdoor data collection is illustrated in Fig. 1.

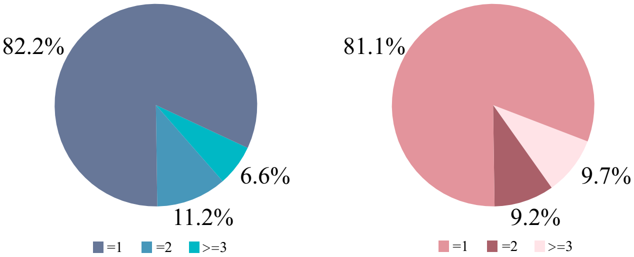

Figure 2. Distribution of the event counts for each activated pixel (has triggered event) within 10 ms window, the left pie chart represents the ON event, the right pie chart represents the OFF event.

1.2. Temporal Window Setting

In the DED framework, we perform a synchronous grid process on the dual-sampled event streams in the time dimension. The shorter the temporal window Δt, the less completeness of scene information; conversely, the longer the Δt, the greater the blend degree of signal and noise events. We conducted a statistical analysis on the raw event within randomly selected 1 s segments of all sequences in the LED dataset. As shown in Fig. 2, it can be found that when the temporal window is 10 ms, most of the activated pixels only have a single event during this period, which means that a large number of noise events are independent on the corresponding pixels without obvious blend situation of signal and noise events. Therefore, to balance the completeness representation of scene information and time resolution requirement, we select a temporal window of $\Delta t =$ 10 ms for binary event frame in the DED framework.

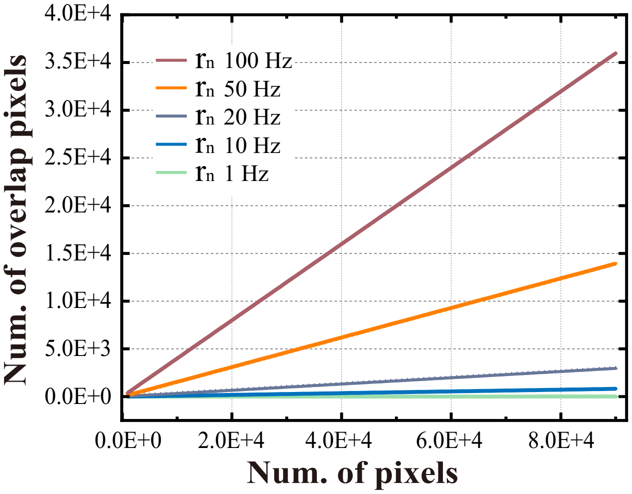

Figure 3. Illustration of the noise event coincidence theoretical measurement with various noise event rates.

1.3. Noise Coincidence Probability Computation

The proposed DED framework is mainly modeled on noise inconsistencies in dual event streams. Here, we explain the probability of simultaneous BA noise occurrence in spatiotemporal synchronized dual-samplings to theoretically support the effectiveness of the DED Framework. Due to the inconsistency of event states (presence or absence) in dual-sampled event data, event generation can be simplified into two situations. Referring to [6], we calculate the probability of noise event observation within a given time window based on the assumption of the Poisson process.

We consider each pixel creates noise event with Poisson rate r_n under pure noise output environment, then the probability p_0 of no event during the temporal window Δt:

$$p_0 = e^{-r_n \Delta t} \tag{1}$$

and the probability p_{1+} of noise event generation in Δt is

$$p_{1+} = 1 - e^{-r_n \Delta t} \tag{2}$$

Further, we can obtain the probability p_s of noise event occurring at the same pixel in the spatiotemporal synchronized dual-sampling:

$$p_s = p_{1+}^{\boldsymbol{x}_{1\#}} \cdot p_{1+}^{\boldsymbol{x}_{2\#}} \tag{3}$$

where $\boldsymbol{x}$ is the coordinate of pixels, $1\#$ and $2\#$ represent the two cameras respectively.

According to the mode where each pixel in the event cameras works independently, we investigate the number of overlapping noise events under different noise rates r_n and pixels quantity N_p. Here, we treat the N_p pixels as independent repeated experiments to obtain the expected values under different conditions, as shown in Fig. 3. Clearly, the noise event trigger rate has the greatest impact on the noise inconsistency property. As the r_n increases, the probability of overlap noise pixels within the observation time window also increases. However, it can be found that this counteracting effect is only significant when the r_n globally beyond 50 Hz. Fortunately, for realistic event cameras, it is rare to have such high global noise event frequency in the temporal dimension, except for a few so-called 'hot pixels'. Thus, the DED framework is generally effective since the average noise event frequency that are typically below 10 Hz.

Layer	Spiking Neuron Num	Neuron Type	Spike Firing Rate
Head	$32 \times H \times W$	LIF	0.127
Down1	$64 \times H \times W$	LIF	0.238
Down2	$128 \times H \times W$	LIF	0.108
Down3	$256 \times H \times W$	LIF	0.058
Res1	$256 \times H \times W$	LIF	0.086
Res2	$256 \times H \times W$	LIF	0.063
Up1	$128 \times H \times W$	LIF	0.037
Up2	$64 \times H \times W$	LIF	0.042
Up3	$32 \times H \times W$	LIF	0.243
EDB-Pred	$1 \times H \times W$	LIF	0.216
DTB-Pred	$1 \times H \times W$	MP	-

Table 1. The DTSNN Architecture Details.

2. Network Details

2.1. Network Architecture

The two branches of DTB and EDB in the proposed DTSNN are both derived from the U-shape model, sharing the same architecture consisting of a head, encoder, residual layer, decoder, and prediction layer, as shown in Table 1. The average spike firing rate of each layer on LED dataset also denotes well preservation of event sparsity and low power consumption potential. The event stream is firstly transformed into continuous binary event frames within a temporal window of Δt (10 ms). For each time step, a $1 \times W \times H$ event binary frame is fed in to the head of DTB and EDB branch respectively, followed by three encoder layers, two residual layers, three decoder layers, and final prediction layer. The difference is that the spiking neurons in the prediction layer of the former are membrane potential neurons (MP neurons) [10], while all neurons in the latter are LIF neurons. Compared to LIF neurons, MP neurons have similar dynamics, which not involves firing process, releasing membrane potential instead of spikes. Thus, the discrete form of MP neurons dynamics can be written as:

$$\begin{cases} V^{n,t} = (1 - \dfrac{1}{\tau})V^{n,t-1} + \dfrac{1}{\tau}U^{n,t} \\ O^{n,t} = V^{n,t} \end{cases} \tag{4}$$

where $O^{n,t}$ denotes the output of the neuron at t in layer n. If we set the membrane time constant τ to 2, a more simplified form is $V^{n,t} = \frac{1}{\tau}(V^{n,t-1} + U^{n,t})$. In our denoising network implementation, the MP neurons layer is used to regress threshold prediction values in the range of 0~1.

2.2. Threshold Map Labeling

Because event cameras capture relative intensity changes, the signal may exhibit local discontinuous arrangements within some certain time interval, such as lower radial relative motion velocity resulting in sparser events trigger. Therefore, it is more difficult to preserve the weaker features of such signal events. Taking this situation into consideration, the labeling of the spiking neurons' threshold map is generated through a longer period of signal events to achieve a denser representation. By accumulating signal events from adjacent temporal windows after denoising through the DED framework, a spatiotemporal-aware map can be obtained to help locate the latent signal or noise region based on spatial-temporal density matrix D. The spatial-temporal density processing is expressed as follows:

$$D_{i,j} = \sum \delta(x,y,t), \quad \substack{(m-1)\Delta t \leq t \leq (m+1)\Delta t \\ i-\frac{L-1}{2} \leq x \leq i+\frac{L-1}{2} \\ j-\frac{L-1}{2} \leq y \leq j+\frac{L-1}{2}} \qquad (5)$$

where $D_{i,j}$ is the density matrix element, L is the odd spatial neighborhood size in the central coordinate (i,j) and m is index of current temporal window. The Dirac impulse function $\delta(x,y,t)$ can be expressed as follows :

$$\delta(x,y,t) = \begin{cases} 1, \; if \, there \, is \, an \, event \; e(x,y,t) \\ 0, \; otherwise \end{cases} \qquad (6)$$

The constructed density matrix D can provide an indication for signal event location in the current spatiotemporal domain, that is, the larger the spatiotemporal density value, the more significant the signal event region. Based on this, a threshold map label F_{th} is further generated as follows:

$$F_{th} = (\beta - \alpha) * \mathcal{N}(\frac{1}{D}) + \alpha \qquad (7)$$

where $\mathcal{N}(\cdot)$ is an operation to normalize the reciprocal of non-negative matrix D to 0~1, α and β are the corresponding threshold intervals. The signal events, the other non-zero values, and the remaining zero values are manually assigned threshold intervals from low to high for signal regions with different spatiotemporal densities.

3. Additional Results of DED

3.1. Quantitative Results

A comprehensive evaluation study on LED dataset are conducted in the manuscript as shown in Fig. 6. Furthermore, based on the evaluation, we conducted quantitative analysis on the following self-proposed metrics: $Retention$, which means the ratio of the number of events after denoising to the original count; $Sparsity$, which indicates the ratio of blank event patches to the total number of patches, measuring the sparsity of events in the spatiotemporal dimension; $EDQE$, which signifies the events spatial distribution quality of events and is defined as:

$$EDQE = Retention * Sparsity \qquad (8)$$

	Raw	Knoise	DWF	STDF	TS	EvFlow	**DED**
Retention	1	0.23	0.61	0.64	0.69	0.73	**0.78**
Sparsity↑	0.12	0.58	0.41	0.35	0.44	0.47	**0.64**
EDQE↑	0.12	0.13	0.25	0.22	0.30	0.34	**0.49**
NIQE↓	4.20	4.41	4.17	4.01	3.95	3.89	**3.53**

Table 2. Quantitative study of the DED framework effectiveness.

The quantitative results are listed in Table 2, where the GT generated by DED is the best in these indirect indices.

3.2. Qualitative Results

Additional results of the paired LED dataset are shown in Fig. 4. Each three rows is the raw event, the denoised event, and the residual noise in turn; the columns from left to right represent adjacent consecutive segments. It can be seen that the LED dataset produced by the DED framework achieves clear separation between signal and noise events. By contrast, the other well-known real-world paired dataset DVSNOISE20 [1] where there are erroneously removed apparent signal events within the residual noise layer as shown in Fig. 5 (in the same way like Fig. 4 for illustration), demonstrating that utilizing multi-modal frame image information to assist in label generation would restrict denoising accuracy increase because data mismatch to some extent.

4. Additional Results of DTSNN

4.1. Results on LED Dataset

Apart from a few methods with closed-source code, we conducted a comparison among different typical methods as widely as possible. The quantitative results on our LED dataset are presented in Table 2 of the manuscript. Traditional event denoising algorithms, such as Knoise [7] and DWF [6], are hardware-friendly and aimed at obtaining a small number of useful signal events with low computational cost, thus exhibiting the lowest performance in terms of signal retain. Other handcrafted algorithms also struggle to achieve higher denoising accuracy, encountering a trade-off where they sacrifice signal preservation in order to achieve a high level of noise removal. Additional results of comparative methods on LED are shown in Fig. 6. Our models perform well on various scenes.

4.2. Results on Other Datasets

Due to the scarcity of real-world paired denoising datasets, we conducted qualitative testing of our model on other publicly available datasets including E-MLB [2], DVSNOISE20 [1] and DSEC [5], and the event cameras they used are different from the cameras used for LED. As shown in Fig. 7, our proposed model still outperforms other methods in denoising effectiveness across different datasets, demonstrating the superiority of our dataset and model.

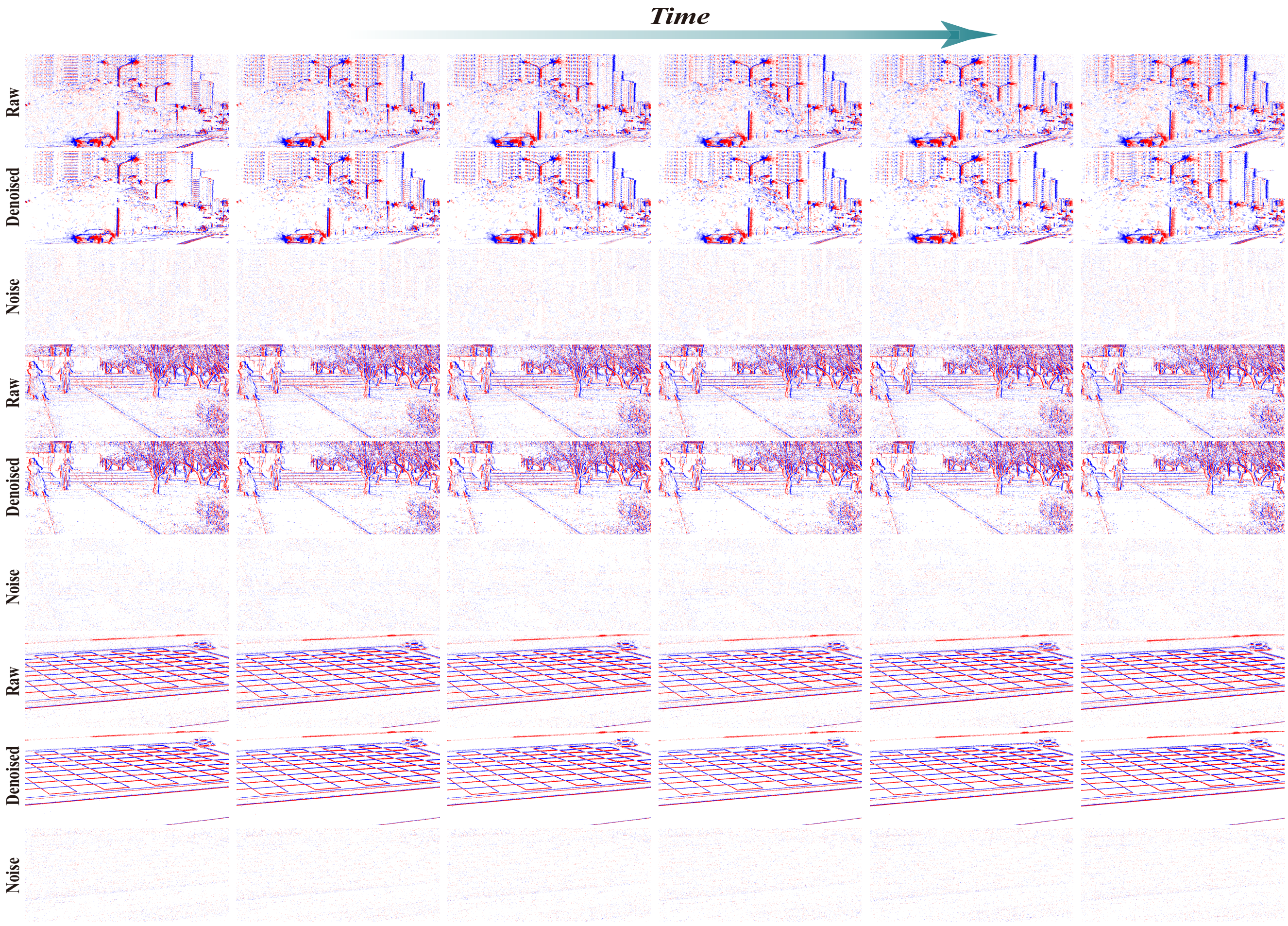

Figure 4. Qualitative results of our DED framework for paired LED dataset generation

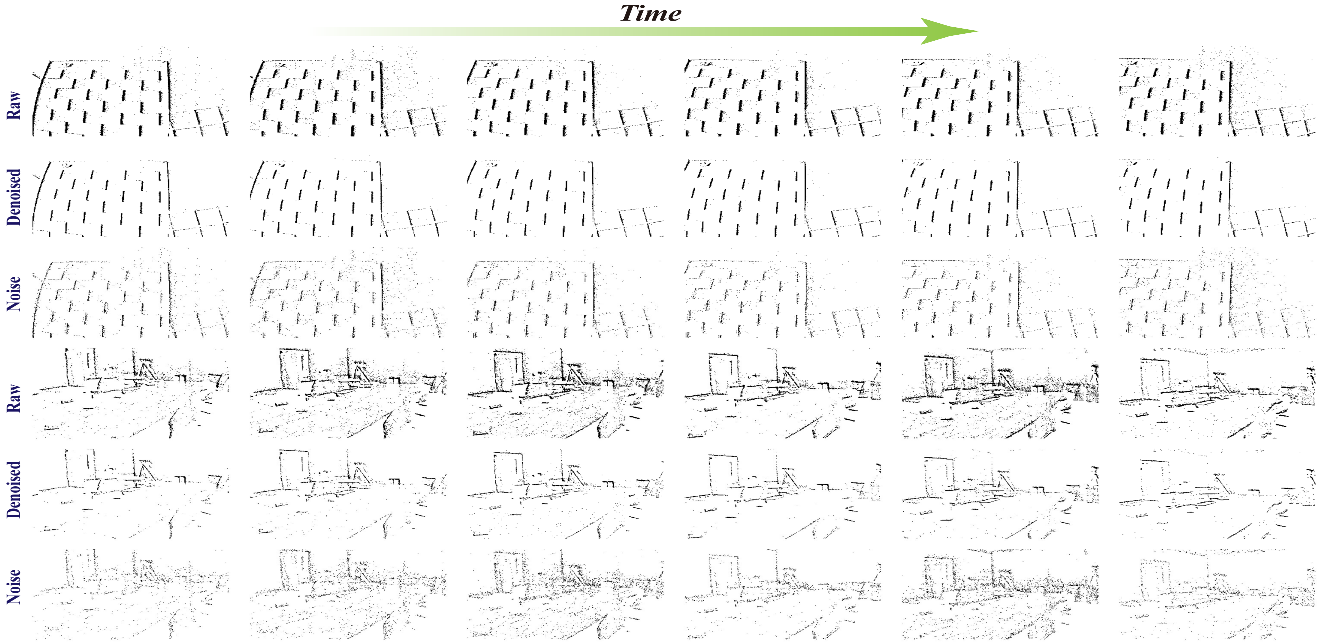

Figure 5. Qualitative results of realistic paired DVSNOISE20 dataset with multi-mode information generation

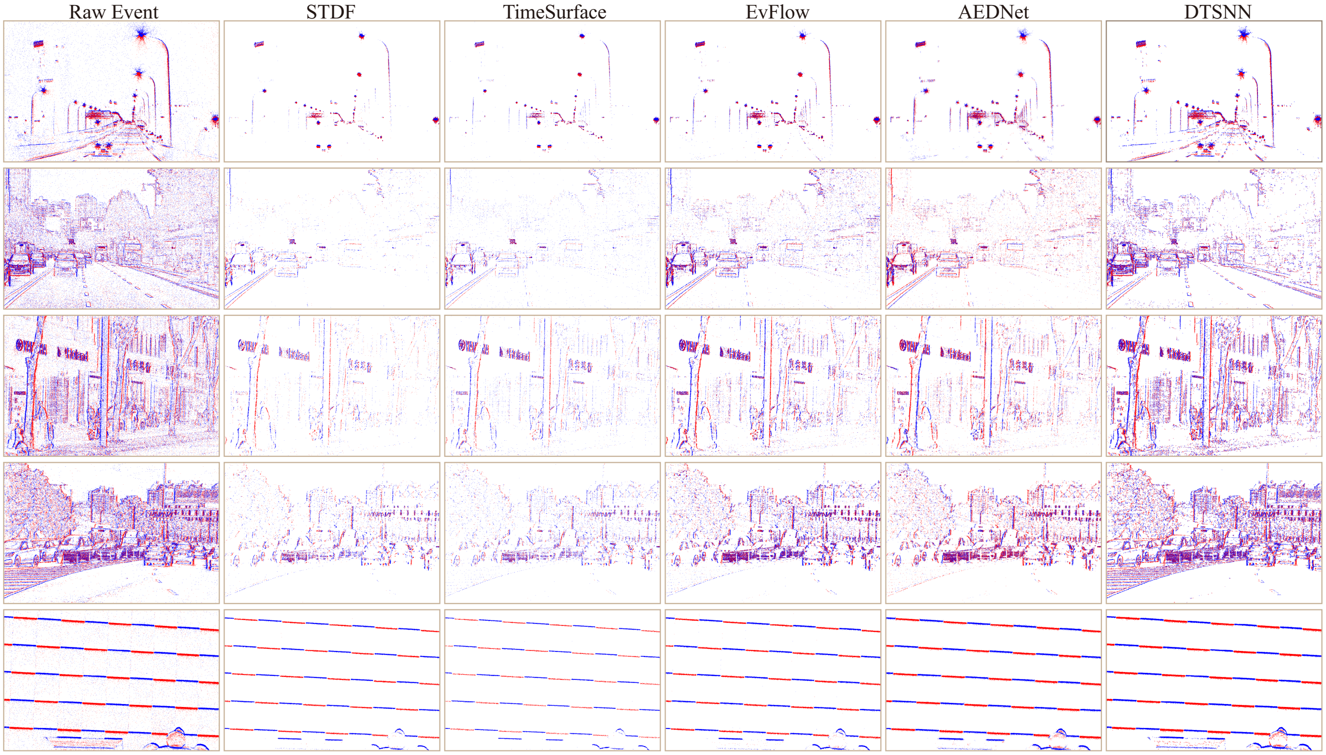

Figure 6. Qualitative denoised results of comparative methods on LED dataset. We recommend zooming in the figure on PC for better visualization.

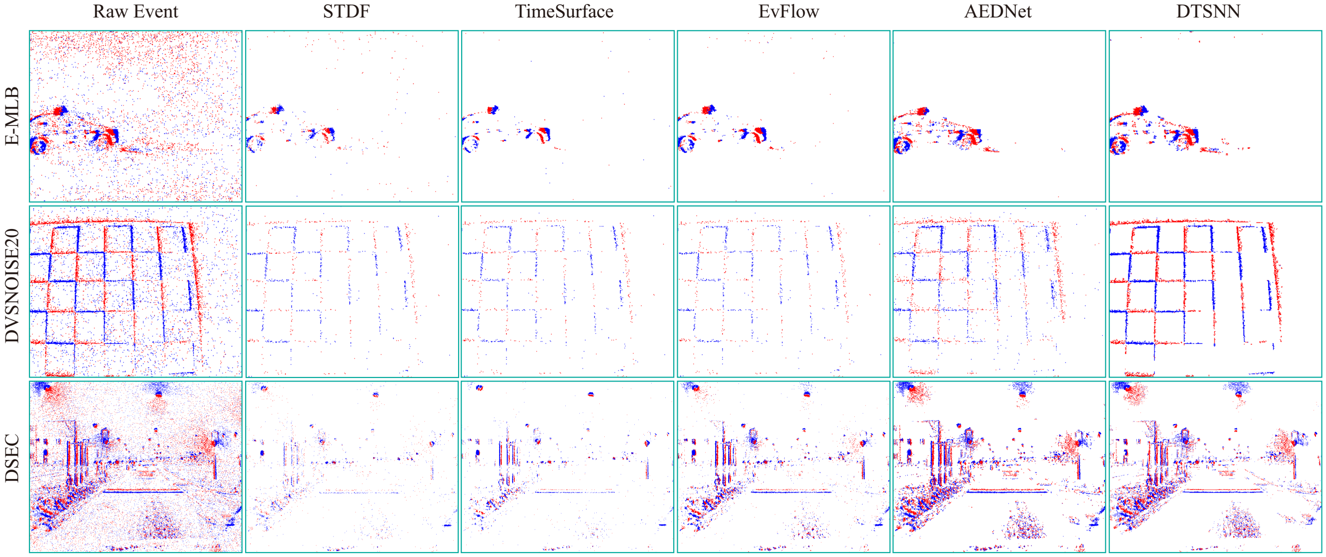

Figure 7. Qualitative denoised results of comparative methods on other public datasets. We recommend zooming in the figure on PC for better visualization.

References

[1] R Baldwin, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2020. 3

[2] Saizhe Ding, Jinze Chen, Yang Wang, Yu Kang, Weiguo Song, Jie Cheng, and Yang Cao. E-mlb: Multilevel benchmark for event-based camera denoising. *IEEE Transactions on Multimedia*, 2023. 3

[3] Huachen Fang, Jinjian Wu, Leida Li, Junhui Hou, Weisheng Dong, and Guangming Shi. Aednet: Asynchronous event denoising with spatial-temporal correlation among irregular data. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1427–1435, 2022.

[4] Yang Feng, Hengyi Lv, Hailong Liu, Yisa Zhang, Yuyao Xiao, and Chengshan Han. Event density based denoising method for dynamic vision sensor. *Applied Sciences*, 10(6): 2024, 2020.

[5] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. 3

[6] Shasha Guo and Tobi Delbruck. Low cost and latency event camera background activity denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):785–795, 2022. 2, 3

[7] Alireza Khodamoradi and Ryan Kastner. $o(n)$-space spatiotemporal filter for reducing noise in neuromorphic vision sensors. *IEEE Transactions on Emerging Topics in Computing*, 9(1):15–23, 2018. 3

[8] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016.

[9] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 1

[10] Beck Strohmer, Rasmus Karnøe Stagsted, Poramate Manoonpong, and Leon Bonde Larsen. Integrating non-spiking interneurons in spiking neural networks. *Frontiers in neuroscience*, 15:633945, 2021. 2

[11] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6358–6367, 2019.