EDformer: Transformer-Based Event Denoising Across Varied Noise Levels

Bin Jiang[1⋆], Bo Xiong[2⋆], Bohan Qu[1], M. Salman Asif[3], You Zhou[4†], and Zhan Ma[1†]

[1] School of Electronic Science and Engineering, Nanjing University
[2] School of Computer Science, Peking University
[3] Department of Electrical and Computer Engineering, UC Riverside
[4] Medical School, Nanjing University
{mazhan,zhouyou}@nju.edu.cn

Abstract. Currently, there is relatively limited research on the background activity noise of event cameras in different brightness conditions, and the relevant real-world datasets are extremely scarce. This limitation contributes to the lack of robustness in existing event denoising algorithms when applied in practical scenarios. This paper addresses this gap by collecting and analyzing background activity noise from the DAVIS346 event camera under different illumination conditions. We introduce the first real-world event denoising dataset, ED24, encompassing 21 noise levels and noise annotations. Furthermore, we propose EDformer, an innovative event-by-event denoising model based on transformer. This model excels in event denoising by learning the spatiotemporal correlations among events across varied noise levels. In comparison to existing denoising algorithms, the proposed EDformer achieves state-of-the-art performance in denoising accuracy, including open-source datasets and datasets captured in practical scenarios with low-light intensity requirements such as zebrafish blood vessels imaging.

Keywords: Event camera · Background activity noise · Denoising · Spatiotemporal correlation

1 Introduction

Event cameras mimic human visual perception by asynchronously outputting events to capture scene motion or brightness changes, providing superior temporal resolution, lower power consumption, and a broader dynamic range compared to traditional cameras, rendering them well-suited for high-speed and high-dynamic-range scenarios [12].

The output of the event camera often incorporates various types of noise, primarily including photon shot noise, dark current shot noise, leakage current noise, and hot pixel noise [14]. Photon shot noise [21] arises from the quantum

⋆ Equally to this work.
† Corresponding author.

nature of photons, while dark current shot noise [27] results from the impact of randomly drifting dark currents in low-light conditions on changes in pixel circuit voltages. Leakage current noise [26] is attributed to the influence of temperature variations and parasitic photocurrents on pixel circuitry, and hot pixel noise [18] is caused by reset switches with unusually low thresholds or exceptionally high dark currents. These four types of noise are collectively denoted as Background Activity (BA) noise. Even in the absence of any scene motion and variations in brightness, these noise events persist, classifying them as signal-independent, i.e., non-informative noise [15]. In dim lighting conditions, BA noise dominates the output of the event camera, significantly impacting the imaging quality.

Numerous efforts have been devoted to BA noise removal, categorized based on denoising approaches into time interval-based, event density-based, image filtering-based, optical flow-based, and learning-based methods. Time interval-based methods [7,16,19] distinguish noise by utilizing the time intervals between triggered events. Event density-based methods [6, 11, 15, 20, 34, 35] differentiate noise by leveraging event density within a specified spatiotemporal range or by considering spatiotemporal distance. Image filtering-based methods [1, 5, 28] transform event sequences into images and subsequently apply image filtering. Optical flow-based methods [9, 24, 25, 32, 33] utilize the motion continuity of events to discern noise. Learning-based methods [2, 3, 10], on the other hand, differentiate noise by learning the features of events.

However, time interval-based and event density-based methods heavily rely on manually crafted threshold parameters, rendering them incapable of adaptively denoising event signals with varying noise rates. Image filtering-based methods can only remove noise at the image level, failing to capture the denoised original event sequence. Optical flow-based methods incur high computational complexity, and optical flow estimation is susceptible to the influence of outliers. Learning-based methods currently transform event sequences into time surfaces or event images for model training. Compared to traditional algorithms that operate event-wise, these learning-based methods sacrifice the temporal granularity of the original event sequence. Additionally, due to the scarcity of denoising datasets, learning-based methods exhibit unstable denoising performance when confronted with event signals from different scenes and noise rates.

In addressing the aforementioned challenges in existing denoising methods, our research motivations are twofold. First, we aim for the proposed denoising model to exhibit generalization across varied BA noise rates. Second, we aim to denoise the raw event sequence directly, preserving its full spatiotemporal granularity for optimal denoising accuracy. In pursuit of these objectives, as shown in Fig. 1, we present the annotated denoising dataset ED24 and introduce an innovative transformer-based event denoising model called EDformer. The primary contributions of our work are as follows:

- We capture pure BA noise from a DAVIS346 event camera using optical instruments, conduct statistical analyses under various illumination conditions, and introduce the first annotated real-world denoising dataset, ED24,

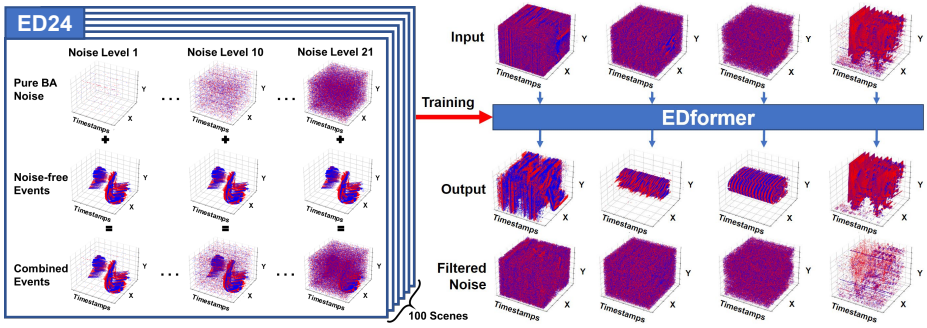


Fig. 1: The overall schematic of our work. On the left side, the ED24 dataset is presented, encompassing 100 distinct scenes, each featuring 21 different noise levels along with noise annotations, suitable for event-wise denoising model training. On the right side, the denoising process of event sequences across various noise levels is depicted using our proposed EDformer model. Red/blue indicates positive/negative polarity event.

comprising 21 noise levels. This dataset effectively addresses the inadequacies in current event denoising training sets.

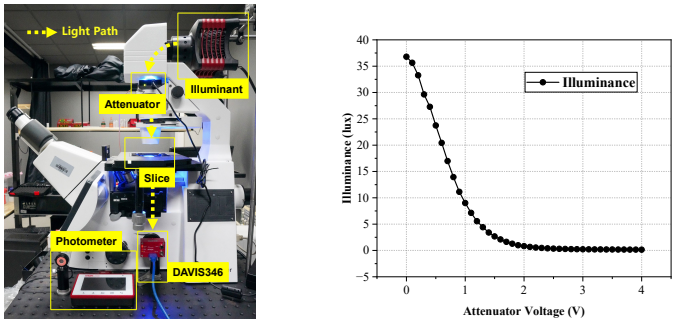
- We introduce the EDformer, a transformer-based denoising model that classifies events for denoising by learning spatiotemporal correlations on an event-by-event basis. In the experimental section, we compared the EDformer with other denoising methods using evaluation metrics such as AUC and MESR. Additionally, we conducted visualization comparisons in the microscopy scene. The experimental results demonstrate the superior performance of our EDformer in event denoising.

2 Noise Analysis

To design effective BA denoising algorithms, it is crucial to understand the genuine characteristics of BA noise. Previous work [15] treated BA noise as a fixed-pattern noise, implying that in relatively low-light conditions (without specifying a particular illuminance), BA noise approximately follows a log-normal distribution. However, they still have two unresolved issues persist: 1) How does BA noise distribute under varying illumination conditions? 2) How should multiple types of noise be modeled when they coexist under certain illumination conditions?

2.1 Statistical Modeling

In order to investigate the two aforementioned issues, as shown in Fig. 2, we utilized an inverted microscope [31] and a DAVIS346 event camera [23] to capture pure BA noise under different illumination conditions. The light source of the inverted microscope maintained a constant intensity. We adjusted the illumination by adding an attenuator in front of the light source and controlling the voltage of the attenuator. The illumination at different voltages was measured using a



(a) Pure BA noise acquisition device **(b)** Attenuator voltage vs. illuminance

Fig. 2: Collecting BA noise by quantitatively adjusting the illuminance.

photometer. The collection process used voltage increments of 0.1, ranging from 0.0 to 4.0 in the attenuator's voltage control. This voltage range corresponds to an illumination range of 36.79 to 0.15 lux, where higher voltage indicates lower illumination. Due to the constant intensity of the light source, the absence of external light interference, and the stationary state of DAVIS346, the collected events are pure BA noise.

The partial results of our statistical modeling of BA noise under varying illumination conditions are presented in Fig. 3, with the complete results available in the Appendix E. This involved computing time intervals between consecutive triggering events at each pixel position and subsequently transforming them into logarithmic frequencies. In the first column of Fig. 3, the attenuator voltages is 0.0V, corresponding to illuminances of 36.79 lux. In this context, the low-frequency range (frequency less than 1Hz) is primarily occupied by positive polarity events. As per [22], these positive polarity events are identified as leakage current noise, displaying a right-skewed log-normal distribution. For the rest column of Fig. 3, the attenuator voltage spans from 1.1V to 2.5V, corresponding to illuminances ranging from 7.13 lux to 0.35 lux. Across this range, decreasing illumination leads to the gradual emergence of mid-frequency noise (frequency greater than 1Hz and less than 10Hz), featuring both positive and negative polarity events. These events exhibit an increasing mean frequency with decreasing illumination, ultimately prevailing in extremely low-light conditions. This suggests that these noise events are attributed to dark current shot noise [27], following a log-normal distribution. Hot pixel noise persists under any illumination condition, primarily in the high-frequency range (frequency greater than 10Hz), and the quantity of hot pixel noise increases gradually as illumination decreases. Due to the diversity and complexity of noise components, the Gaussian mixture fitting results in Fig. 3 only provide a rough illustration of the components of BA noise in different frequency ranges and their variations with decreasing illumination. Analyzing the changes in BA noise reveals that the BA noise at any illumination level is composed of various noise components, with a frequency span extending across four orders of magnitude.

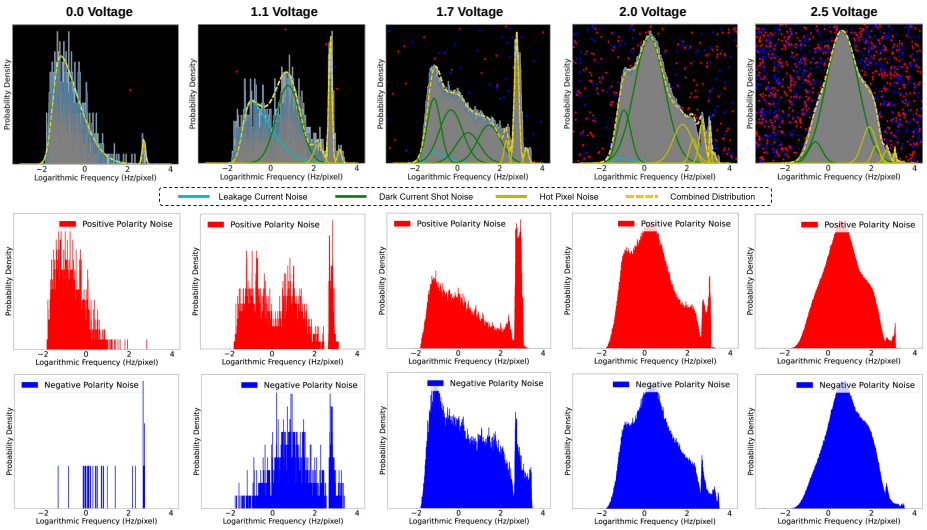


Fig. 3: The first row depicts the log-frequency probability density of all BA noise across five illumination conditions, fitted with the mixture Gaussian distribution. The black background displays the pure BA noise collected within 33 ms, where the red point indicates positive polarity noise, and the blue point indicates negative polarity noise. The second and third row represent the log-frequency probability density of positive and negative polarity BA noise, respectively.

Table 1: The comparison of public event denoising datasets

Datasets	Camera	Resolution	APS	IMU	Scenes	Sequences	Capture/s	DoF	Noise Level	Noise Label
DVSNOISE20 [3]	DAVIS 346	346 ×260	Gray	✓	16	48	807	Cam.	-	-
RGBDAVIS [9]	DAVIS 240	190 ×180	RGB	✓	20	20	122	All.	-	-
ENFS [10]	DAVIS 346	224 ×125	-	-	1	100	4238	Obj.	-	-
DND21 [15]	DAVIS 346	346 ×260	-	-	-	8	-	All.	2	✓
E-MLB [8]	DAVIS 346	346 ×260	Gray	✓	100	1200	7300	All.	4	-
ED24 (ours)	DAVIS 346	346 ×260	-	✓	100	**2100**	7300	All.	**21**	✓

2.2 Dataset Creation

Previous work [15] used a log-normal distribution to model BA noise. However, in reality, the distribution of BA noise varies significantly under different illumination conditions and cannot be adequately simulated with a single distribution parameter. Given the disparity between simulating noise and real-world noise, we directly incorporated the collected pure BA noise into the noise-free event sequences, creating the ED24 dataset required for denoising model training.

Specifically, we employed an inverted microscope and a DAVIS346 event camera to respectively capture pure BA noise under 21 different illumination conditions, ranging from 1.5V to 3.5V for attenuator voltage (BA noise below 1.5V was too sparse, and noise distribution above 3.5V was nearly uniform). Each illumination condition was recorded for one minute. Subsequently, we randomly

sampled BA noise of corresponding duration based on the timestamps of noise-free event sequences. The sampled noise was then combined with the noise-free events, and the combined events were reordered based on timestamps. It is noteworthy that the noise-free event sequences were generated using the DAVIS346 event camera capturing 100 indoor and outdoor scenes under well-lit conditions. Due to the ample brightness, BA noise had low frequency and was extremely sparse, enabling its complete removal using a straightforward BAF algorithm [7]. In our experiments, we set the BAF's time interval threshold to 1ms. As shown in Tab. 1, our ED24 dataset stands out as the first annotated real-world event denoising dataset, encompassing a range of 21 noise levels. For additional ED24 dataset details, please refer to the Appendix D.

In practical terms, our approach to constructing the denoising dataset ED24 has two potential limitations: 1) Owing to variations in the number of photons per unit time, valid events in low-light conditions tend to be sparser compared to those in well-lit conditions. Consequently, directly merging noise-free data collected in well-lit conditions with BA noise in low-light conditions may not accurately reflect real-world scenarios. 2) Our noise capture was limited to the DAVIS346 under different brightness conditions, and variations may exist across different sensors. To address these issues, our future work will explore the relationship between brightness and the sparsity of valid events. Additionally, we plan to integrate noise patterns from other event cameras into the dataset. Nevertheless, the current ED24 dataset is already sufficient for neural network models to learn the spatiotemporal differences in correlation between valid events and BA noise. The specific denoising performance will be discussed and analyzed in the experimental section.

3 Denoising Model

3.1 Problem Definition

According to the pixel circuitry of the event camera [22, 30], when the light signal L is incident upon the photo-sensor, it is converted into the current $I = I_p + I_{dark}(I_p \propto L)$, where I_p is the the photocurrent and I_{dark} is the dark current. Subsequently, current I undergoes logarithmic transformation through a feedback diode, resulting in the voltage V_p. Later, it is further amplified into the voltage change $\Delta V_d(t)$. The voltage change value is also influenced by the unavoidable junction leakage current. When the voltage change $\Delta V_d(t)$ reaches the positive threshold $-\theta_{ON}$ or the negative threshold θ_{OFF}, an event is triggered:

$$\begin{cases} \Delta V_d(t) \leq -\theta_{ON} & \text{ON event} \\ \Delta V_d(t) \geq \theta_{OFF} & \text{OFF event} \\ -\theta_{ON} < \Delta V_d(t) < \theta_{OFF} & \text{No event} \end{cases} \tag{1}$$

This process results in a sequence of N-event formally denoted as $\mathbf{E} = \{\boldsymbol{e}_i\}_{i=1}^{N}$. Each event $\boldsymbol{e}_i = \{\boldsymbol{u}_i, p_i, t_i\}$ is a tuple that records pixel position $\boldsymbol{u}_i = (x_i, y_i)$,

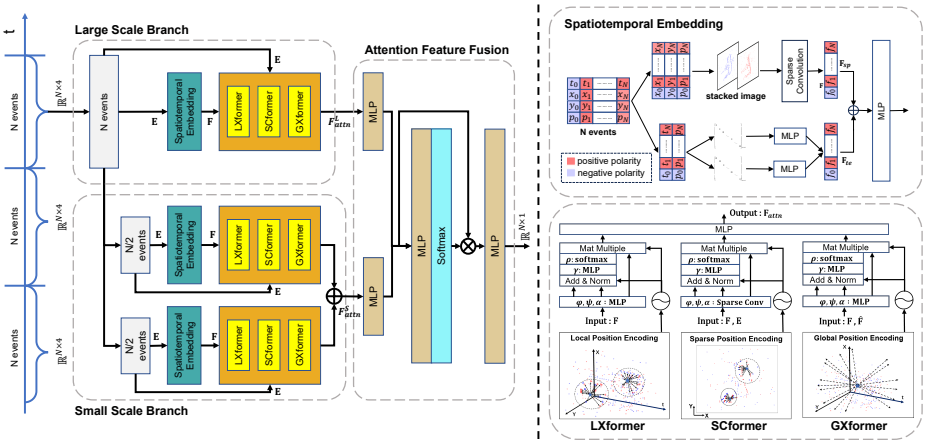


Fig. 4: Left side illustrates the pipeline of our EDformer, featuring the large/small-scale branch and attention feature fusion module. Each branch comprises spatiotemporal embedding and three-way attention modules. The attention features $\mathbf{F}^{L}_{attn}$ and $\mathbf{F}^{S}_{attn}$ from large/small-scale branches are fused using the attention feature fusion module to obtain the event classification result $\mathbf{F}_{out}$. Right side presents the key computational steps in the spatiotemporal embedding and three-way attention modules.

where the event is located, polarity $p_i \in \{-1, 1\}$ and timestamp t_i. p_i indicates the increase or decrease of pixel brightness.

Due to the fact that event cameras output both valid signals and noise in the form of events, direct differentiation is highly challenging. However, the generation of valid events is generally linked to changes in brightness caused by object motion, characterized by dense and continuous patterns. On the other hand, the generation of BA noise is random, and in the spatiotemporal dimension, it tends to be more sparse and irregular compared to valid events. In other words, the distinction between valid events and BA noise lies in their spatiotemporal characteristics. Therefore, our proposed EDformer distinguishes valid events from fixed-pattern noise by calculating spatiotemporal correlations on an event-by-event basis, treating the denoising task as an event classification problem.

3.2 Network Architecture

As depicted in Fig. 4, the EDformer segments the entire event sequence into multiple segments based on timestamp order, with each segment containing N events, and provides binary classification outcomes for each of them. The model comprises three main modules: the large-scale branch, the small-scale branch, and the attention feature fusion. The large-scale and small-scale branches are responsible for event-wise feature extraction and spatiotemporal correlation calculation across different temporal scales in the event sequence. The attention feature fusion module is designed to integrate features from events across various temporal scales, ultimately generating the final classification results.

Spatiotemporal Embedding For N input events $\mathbf{E} = \{\boldsymbol{e}_i\}_{i=1}^{N}$, the spatiotemporal embedding module performs feature extraction in both spatial and temporal dimensions based on the events' pixel coordinates and timestamp information. In the spatial dimension, the N events are stacked into positive and negative polarity channels of an event image, based on their polarities and pixel coordinates. These channels record the quantity of positive and negative polarity events at each pixel position. Subsequently, considering the sparsity of events, we employ sparse convolution [13] to extract spatial features at pixel positions, which yields N spatial embedding vectors $\mathbf{F}_{sp}$. The convolution is computed only when the pixel is at the center of the convolution kernel, and events stacked at the same pixel location share the convolution results. In the temporal dimension, the N events undergo filtering for positive and negative polarity using two diagonal matrices. Then, the timestamps of positive and negative polarity events are separately input into two different MLPs to compute the temporal embedding tensors $\mathbf{F}_{te}$ for N events. Consequently, we concatenate $\mathbf{F}_{sp}$ and $\mathbf{F}_{te}$, and downscale them to a spatiotemporal embedded vectors $\mathbf{F} = \mathrm{MLP}(\mathrm{Concat}(\mathbf{F}_{sp}, \mathbf{F}_{te}))$, where the index of spatiotemporal embedded vector $\boldsymbol{f}_i$ correspond one-to-one with that of the event $\boldsymbol{e}_i$.

Event-wise Position Encoding Similar to point clouds, event sequences can also be construed as ensembles within an irregularly embedded metric space, where the attributes of the constituent elements pertain to spatiotemporal information. H. Zhao et al. [37] introduces the utilization of vector attention to address collections within such irregular embedding metric spaces, establishing the consequential significance of judiciously implemented positional encoding. The mathematical formulation thereof is as follows:

$$\mathbf{attn}_i = \sum_{\boldsymbol{f}_j \in \chi(i)} \rho(\gamma(\varphi(\boldsymbol{f}_i) - \psi(\boldsymbol{f}_j) + \delta)) \odot (\alpha(\boldsymbol{f}_j) + \delta) \tag{2}$$

where $\boldsymbol{f}_i$ is the input embedded vector, and $\odot$ represents Hadamard product. The subset $\chi(i) \subseteq \chi$ is a set of vectors in a selected neighborhood of $\boldsymbol{f}_i$. φ, ψ and α are feature transformations, such as sparse convolution or MLPs. ρ is a normalization function such as softmax and γ is a mapping function (e.g., an MLP) that produces attention vectors for feature aggregation. δ is a position encoding function. As our spatiotemporal embedding module has established a one-to-one correspondence between embedding vectors and events based on their indices, in the subsequent computations, $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ in Eq. (2) can be obtained based on the indices of $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$.

According to Eq. (2), we have devised three transformer layers tailored for event sequences. In contrast to the positional encoding methods in [37], which exclusively address local contexts, our proposed LXformer, SCformer, and GXformer concurrently consider both local and global spatiotemporal information for event positional encoding, thereby enhancing the comprehension of the intrinsic structure of event sequences for effective event classification.

Firstly, the continuity of motion establishes close correlations among local events in the spatiotemporal dimensions. To construct spatiotemporal correlations between adjacent events, we employ K-nearest neighbors (KNN) to executing local position encoding in LXfomer:

$$\delta_{ij}^{loc} = \mathrm{MLP}(\sqrt{(x_i - x_j^{loc})^2 + (y_i - y_j^{loc})^2 + (t_i - t_j^{loc})^2}) \tag{3}$$

where x_i, y_i, t_i are the coordinates and timestamp of event $\boldsymbol{e}_i$, and $x_j^{loc}, y_j^{loc}, t_j^{loc}$ are the coordinates and timestamp of event $\boldsymbol{e}_j^{loc}$, which is one of the k_l nearest event to $\boldsymbol{e}_i$. Notably, the computed self-attention weights $\mathbf{attn}_i^{loc}$ computed by LXformer represent the local spatiotemporal correlations between events. This computational approach is analogous to denoising algorithms based on event density, where both methods involve noise removal through the examination of event features within a local spatiotemporal scope.

Subsequently, to further elucidate the XY-dimensional relevance of events, we utilize the ball query algorithm [29] to retrieve k_{sc} events $\{\boldsymbol{e}_j^{sc}\}_{j=1}^{k_{sc}}$ within the $m \times m$ local window of each event $\boldsymbol{e}_i$ to executing sparse convolutional positional encoding in SCfomer:

$$\delta_{ij}^{sc} = \mathrm{MLP}(\sqrt{(x_i - x_j^{sc})^2 + (y_i - y_j^{sc})^2}) \tag{4}$$

where x_j^{sc}, y_j^{sc} are the coordinates of event $\boldsymbol{e}_j^{sc}$. The self-attention weights $\mathbf{attn}_i^{sc}$ computed by SCformer signify the local spatial correlations between events in the XY dimension, which bears striking resemblance to denoising methods based on image filtering.

The overarching global spatiotemporal correlations among events afford a holistic grasp of the motion trends across the entire event sequence, a pivotal consideration for high-level visual tasks. To effectively extract global spatiotemporal correlations, we employ subsampling with a rate of r for down-sampling the farthest events to select representative events $\hat{\mathbf{E}}$, whose corresponding embedded vectors are denoted as $\hat{\mathbf{F}}$, and then utilize KNN to select k_g events $\{\boldsymbol{e}_j^g\}_{j=1}^{k_g} \in \hat{\mathbf{E}}$ to executing global position encoding in GXfomer:

$$\delta_{ij}^{g} = \mathrm{MLP}(\sqrt{(x_i - x_j^{g})^2 + (y_i - y_j^{g})^2 + (t_i - t_j^{g})^2}) \tag{5}$$

where x_j^g, y_j^g, t_j^g are the coordinates and timestamp of event $\boldsymbol{e}_j^g$. Specifically, the embedding vector $\boldsymbol{f}_j$ used in the computation of global attention $\mathbf{attn}_i^g$ is sampled from $\boldsymbol{f}_j = \max_{\boldsymbol{e}_r \in \mathbf{E}_r} (\mathrm{MLP}(\mathrm{Concat}(\boldsymbol{e}_r, \boldsymbol{f}_r)))$, where $\mathbf{E}_r$ is the set of $\lfloor \frac{1}{r} \rfloor$ nearest events of e_g, $\boldsymbol{f}_r$ is the corresponding embeded vector of $\boldsymbol{e}_r$.

Finally, we concatenate local spatiotemporal attention $\mathbf{attn}_i^{loc}$, local spatial attention $\mathbf{attn}_i^{sc}$, and global spatiotemporal attention $\mathbf{attn}_i^g$, and pass them through an MLP to obtain the fused attention of the three transformer layers:

$$\mathbf{attn}_i = \mathrm{MLP}(\mathrm{Concat}(\mathbf{attn}_i^{loc}, \mathbf{attn}_i^{sc}, \mathbf{attn}_i^g)) \tag{6}$$

It is crucial to emphasize that the index of the fused attention $\mathbf{attn}_i$ also correspond one-to-one with that of the event $\boldsymbol{e}_i$, and the attention feature for all events are denoted as $\mathbf{F}_{attn} = \{\mathbf{attn}_i\}_{i=1}^{N}$.

Attention Feature Fusion Considering that even for the same event, the attention feature across the two scale branches may differ, which further impact the classification results $\mathbf{F}_{out}$. So we design a simple and efficient attention feature fusion module to comprehensively consider the same event at different scales:

$$\mathbf{F}_{out} = \text{MLP}(\text{softmax}(\text{MLP}(\mathbf{F}')) \odot \mathbf{F}') \tag{7}$$

where $\mathbf{F}' = \text{MLP}(\mathbf{F}^L_{attn}) + \text{MLP}(\mathbf{F}^S_{attn})$, and $\mathbf{F}^L_{attn}$ and $\mathbf{F}^S_{attn}$ are the attention feature from large-scale and small-scale branch. The concatenated tensor $\mathbf{F}'$ undergoes transformative mapping via an MLP. Subsequently, the softmax function yields a set of weights reflective of the significance of individual elements. These weights are then employed in a weighted summation through tensor multiplication, facilitating the amalgamation of features and accentuating pertinent components in the final output.

4 Experiments

4.1 Datasets and Metrics

Currently, the public datasets for event-based denoising primarily include DVS-NOISE20 [3], RGBDAVIS [9], ENFS [10], DND21 [15], E-MLB [8]. Among them, only DND21 has event-wise noise annotations that can be used for training denoising models based on event classification. However, DND21 has a very small amount of data and is obtained through v2e simulation [17]. Its spatiotemporal distribution differs from real-world event sequences, resulting in the denoising effect of trained models not being well applicable to real data. In response to this, we have created a real-world dataset, ED24, with noise annotations for training our proposed EDformer denoising model. The trained model is then tested for denoising performance on the other public denoising datasets.

The mainstream evaluation metrics for event denoising include MESR [8], RPMD [3], and ROC/AUC [15]. MESR evaluates denoising performance by projecting events into a warped event image and calculating image contrast metric, which can be directly tested on event sequence. RPMD evaluates denoising performance by measuring APS intensity and predicting DVS behavior based on IMU motion. ROC/AUC evaluates denoising accuracy by calculating False Positive Rate (FPR) and True Positive Rate (TPR) using noise-annotated data.

4.2 Experimental Setup

We trained our EDformer on an NVIDIA RTX 3090, utilizing the whole ED24 dataset for training. During training, the batch-size is 96, and $N = 4096$ events were randomly sampled from the training set without any data augmentation, but all events can be input during inference. In the large-scale branch, $k_l = k_{sc} = k_g = 16$, $m = 9$, and $r = 8$. In the small-scale branch, $k_l = k_{sc} = k_g = 16$, $m = 9$, and $r = 16$. The model was trained from scratch for 60 epochs using the AdamW optimizer with a learning rate of 0.001, aiming to minimize the cross-entropy loss [36].

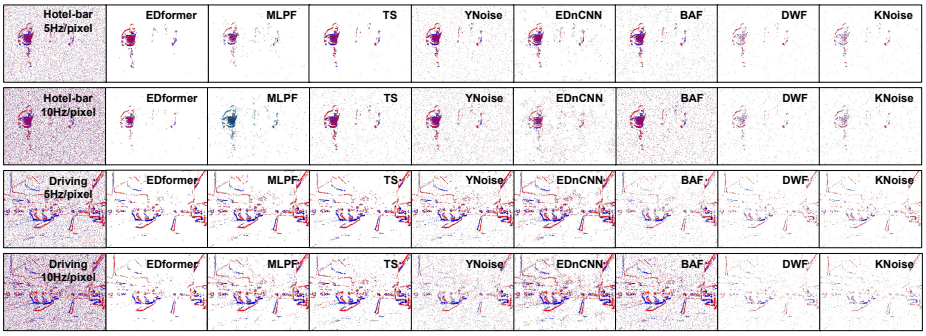


Fig. 5: Visual comparison of DND21 dataset at 5Hz/pixel and 10Hz/pixel noise rates

4.3 Quantitative Evaluation

To accurately assess denoising accuracy, we conducted AUC experiments following the settings in [15]. We employed v2e synthetic shot noise with a frequency range of 1 to 10 Hz/pixel, a COV of 0.5 decades FPN, and added them to the hotel-bar and driving datasets. For BAF, YNoise, TS and KNoise, we swept the correlation time t in $[2, 200]$ ms. For DWF, we swept the distance threshold s with magnitude in $[10, 100]$ pixels. For MLPF, EDnCNN and EDformer, we swept the classification threshold θ from 0 to 1. The AUC results are illustrated in Tab. 2 and the corresponding ROC are shown in the Appendix B. The visual comparison of denoising of the two scenes at 5Hz/pixel and 10Hz/pixel is shown in Fig. 5. It is evident that our proposed EDformer consistently demonstrates superior performance, achieving the highest AUC and showcasing robust generalization in denoising accuracy across varying shot noise rates. Notably, under high BA noise rates, the denoising performance of EDformer surpasses that of other methods, affirming its efficacy in challenging low-light conditions.

Table 2: The AUC of different denoising methods on DND21 datasets at different shot noise rates. The best bolded and the second underlined.

Methods	1 Hz/pixel		3 Hz/pixel		5 Hz/pixel		7 Hz/pixel		10 Hz/pixel	
	Hotel-bar	Driving	Hotel-bar	Driving	Hotel-bar	Driving	Hotel-bar	Driving	Hotel-bar	Driving
KNoise [19]	0.6773	0.6296	0.6521	0.6230	0.6703	0.6235	0.6583	0.6164	0.6413	0.6142
DWF [15]	0.9268	0.7409	0.8930	0.7099	0.8620	0.6901	0.8338	0.6747	0.7958	0.6563
BAF [7]	0.9535	0.8479	0.9197	0.8155	0.8916	0.7930	0.8662	0.7732	0.8366	0.7479
EDnCNN [3]	0.9573	0.8873	0.9371	0.8771	0.9365	0.8748	0.9254	0.8654	0.9006	0.8574
Ynoise [11]	0.9690	<u>0.9409</u>	0.9517	0.9240	0.9234	0.9093	0.9177	0.8972	0.8987	0.8800
TS [20]	<u>0.9716</u>	0.9307	<u>0.9721</u>	<u>0.9260</u>	0.9606	<u>0.9270</u>	0.9654	<u>0.9241</u>	0.9620	<u>0.9202</u>
MLPF [15]	0.9704	0.8887	0.9718	0.8873	<u>0.9704</u>	0.8845	<u>0.9691</u>	0.8817	<u>0.9634</u>	0.8761
EDformer (ours)	**0.9928**	**0.9541**	**0.9891**	**0.9472**	**0.9845**	**0.9424**	**0.9792**	**0.9343**	**0.9699**	**0.9264**

Furthermore, we utilized the inverted microscope and DAVIS346 event camera to capture additional event data of zebrafish blood vessels. Due to the po-

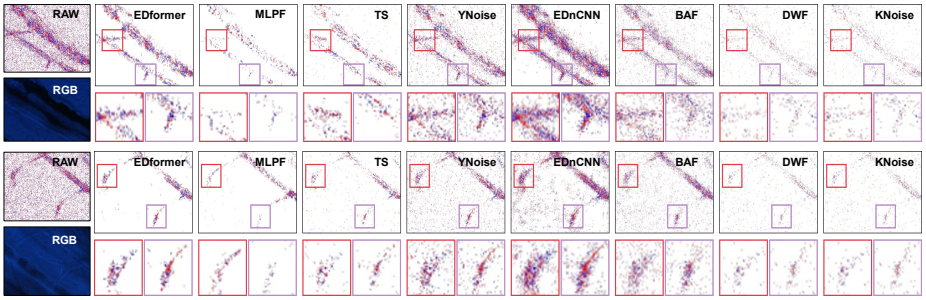


Fig. 6: Observation of zebrafish blood vessels using the inverted microscope and DAVIS346 event camera, comparing denoising performance of different methods. The proposed method effectively remove the BA noise and retain the valid information of raw event sequence.

tential harm posed by intense light to zebrafish and the subsequent rise in the surrounding temperature, causing discomfort to the organisms, the attenuator voltage was set to 3.5V, creating an extremely low-light environment. Inadequate illumination has rendered CMOS chip in DAVIS346 incapable of capturing clear RGB images of zebrafish blood vessels. This underscores the advantage of event camera with high dynamic range. However, the raw event sequence is also plagued by a considerable amount of BA noise, necessitating effective removal. We visually compared the denoising algorithms mentioned in Tab. 2, and the results are shown in Fig. 6. It can be observed that our EDformer is capable of accurately removing BA noise even in low-light microscopic scenarios. Moreover, it retains more complete details of blood vessels compared to other denoising algorithms. This advantage in denoising accuracy significantly expands the future practical application of event camera in microscopy.

In order to further validate the denoising generalization of our proposed EDformer, we conducted MESR testing on E-MLB, RGBDAVIS and DND21 datasets according to the settings described in [8]. As shown in Tab. 3, our model achieves the highest MESR score on RGBDAVIS and ranks second on E-MLB and DND21. It is worth noting that our model achieved the highest MESR on the the E-MLB (Night) ND64 dataset. This further demonstrates the superior denoising performance of EDformer under extremely low-light conditions compared to other denoising methods. During the MESR testing process, we identified a certain limitation of MESR in evaluating event over-denoising. To illustrate this, we conducted additional analysis in the Appendix C.

4.4 Ablation Experiments

To further validate the impact of the three transformer layers on denoising performance, we designed ten ablation experiments as shown in Tab. 4, which illustrates the impact of removing different components on denoising accuracy. Specifically, Exp. #2 removes SCformer and GXformer, Exp. #3 removes GX-

Table 3: The mean ESR (MESR) results of different denoising methods on public available event denoising datasets. The best bolded and the second underlined.

Methods	E-MLB (Daylight)				E-MLB (Night)				RGBDAVIS		DND21
	ND1	ND4	ND16	ND64	ND1	ND4	ND16	ND64	Indoor	Outdoor	-
Raw	0.821	0.824	0.815	0.786	0.89	0.824	0.786	0.768	0.905	0.776	0.869
BAF [7]	0.861	0.869	0.876	0.89	0.946	0.973	0.992	0.942	0.943	0.891	0.92
KNoise [19]	0.846	0.837	0.83	0.807	0.954	0.956	0.871	0.817	0.934	0.895	0.887
DWF [15]	0.878	0.876	0.866	0.865	0.923	0.962	0.988	0.932	0.923	0.89	0.905
EvFlow [32]	0.848	0.878	0.868	0.833	0.969	0.983	0.889	0.797	0.829	1.061	1.006
YNoise [11]	0.866	0.863	0.857	0.821	1.009	0.943	0.875	0.792	0.825	1.077	0.966
TS [20]	0.877	0.887	0.87	0.837	1.033	0.944	0.886	0.797	0.837	1.12	0.985
†IETS [4]	0.772	0.785	0.777	0.753	0.950	0.823	0.804	0.711	0.762	0.988	0.900
†GEF [9]	**1.051**	0.938	0.935	0.927	1.027	0.955	0.946	0.935	<u>1.031</u>	0.986	0.932
MLPF [15]	0.851	0.855	0.846	0.84	0.926	0.928	0.91	0.906	0.983	0.932	0.944
EDnCNN [3]	0.887	0.908	0.903	0.912	1.001	**1.024**	**1.079**	<u>1.086</u>	0.982	1.014	0.977
†EventZoom [10]	<u>0.996</u>	**0.988**	**0.996**	**0.97**	**1.055**	1.007	1.01	0.988	0.93	<u>1.135</u>	**1.059**
EDformer (Ours)	0.952	<u>0.955</u>	<u>0.956</u>	<u>0.942</u>	<u>1.048</u>	<u>1.019</u>	<u>1.076</u>	**1.099**	**1.051**	**1.17**	<u>1.041</u>

† The result is cited from [8], for which the source code has not been released to the public.

former, Exp. #9 removes the small-scale branch, and Exp. #10 removes the large-scale branch. Exp. #1 is the optimal model parameter settings, while the other nine comparative experiments investigate the effects of module combination, spatiotemporal dimension, and large/small-scale branch on the model's denoising performance.

Module Combination When comparing Exp. #2, #3, and #4, adding LXformer, SCformer, and GXformer modules improves the model's classification accuracy. This suggests that EDformer's denoising performance relies on local spatiotemporal correlations, and incorporating calculations for local space and global spatiotemporal correlations further enhances denoising accuracy. In contrast, comparing Exp. #1 and #4 reveals that inhibiting the GXformer in the large-scale branch actually improves denoising accuracy. This indicates that global spatiotemporal correlation may not be suitable for large temporal scale.

Spatiotemporal Dimension When comparing Exp. #5, #6, #7, and #8 as a group, it can be observed that excessively large or small values for KNN in LXformer, and an overly small ball query size in SCformer and a reduction in the global sample number in GXformer both result in a decrease in the model's denoising accuracy.

Large/Small-Scale Branch When comparing Exp. #9 and #10, using a fixed input quantity instead of considering event at different temporal scales significantly reduces the model's denoising performance. This is because the fixed event input quantity may not accurately capture the varying number of events triggered in a scene due to changes in environmental brightness and motion. The design of branches for different temporal scales effectively addresses this issue.

Inference Time Since our attention module only considers k neighbours around each event as discussed in Eq. (2). The computational complexity for attention, therefore, reduces from $O(N^2)$ to $O(Nk)$. We tested with a sequence of length $N = 89960$, and the GPU memory consumption was 4GB. During inference, determining whether each event is BA noise takes approximately 22 μs. If we simplify the model components to only use LXformer, the inference time can be reduced to $\sim 9\mu s$ per event, but the denoising accuracy will decrease.

Table 4: Ablation experiments of EDformer parameters

Exp.	Large Scale Branch			Small Scale Branch			AUC (5 Hz/pixel)	
	k_l	m	r	k_l	m	r	Hotel-bar	Driving
#1	16	9	8	16	9	16	**0.9845**	**0.9424**
#2	16	-	-	16	-	-	0.9720	0.9234
#3	16	9	-	16	9	-	0.9779	0.9364
#4	16	9	16	16	9	16	0.9841	0.9415
#5	8	9	8	8	9	16	0.9808	0.9396
#6	32	9	8	32	9	16	0.9815	0.9324
#7	16	5	8	16	5	16	0.9835	0.9379
#8	16	9	8	16	5	8	0.9836	0.9386
#9	16	9	8	-	-	-	0.9628	0.9051
#10	-	-	-	16	9	16	0.9608	0.8954

5 Conclution

This paper extends the research on the BA noise under various illumination conditions. We introduced ED24, the first annotated real-world event denoising dataset with 21 noise levels, and proposed EDformer, a transformer-based denoising model, which achieves event denoising through event-wise classification by learning the spatiotemporal correlations of events. EDformer outperforms existing methods, showcasing state-of-the-art denoising accuracy. These contributions aim to enhance the understanding of event camera BA noise and provide valuable resources for future research in the event denoising field.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.62022038, 62071219, 62371006). Our code is publicly accessible at `https://github.com/NJUVISION/EDformer`.

References

1. Acharya, J., Caycedo, A.U., Padala, V.R., Sidhu, R.R.S., Orchard, G., Ramesh, B., Basu, A.: Ebbiot: A low-complexity tracking algorithm for surveillance in iovt

using stationary neuromorphic vision sensors. In: 2019 32nd IEEE International System-on-Chip Conference (SOCC). pp. 318–323 (2019). https://doi.org/10.1109/SOCC46988.2019.1570553690
2. Afshar, S., Ralph, N., Xu, Y., Tapson, J., Schaik, A.v., Cohen, G.: Event-based feature extraction using adaptive selection thresholds. Sensors **20**(6), 1600 (2020), https://www.mdpi.com/1424-8220/20/6/1600
3. Baldwin, R.W., Almatrafi, M., Asari, V., Hirakawa, K.: Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. In: CVPR. pp. 1698–1707 (2020). https://doi.org/10.1109/CVPR42600.2020.00177
4. Baldwin, R.W., Almatrafi, M., Kaufman, J.R., Asari, V., Hirakawa, K.: Inceptive event time-surfaces for object classification using neuromorphic cameras. In: Image Analysis and Recognition. pp. 395–403. Springer International Publishing (2019)
5. Bose, S.K., Singla, D., Basu, A.: A 51.3-tops/w, 134.4-gops in-memory binary image filtering in 65-nm cmos. IEEE Journal of Solid-State Circuits **57**(1), 323–335 (2022). https://doi.org/10.1109/JSSC.2021.3098539
6. Chen, Y., Huang, Y., Li, F., Zeng, X., Li, W., Wang, M.: Denoising method for dynamic vision sensor based on two-dimensional event density. In: 2023 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–4 (2023). https://doi.org/10.1109/ISCAS46773.2023.10181865
7. Delbruck, T.: Frame-free dynamic digital vision. In: Proceedings of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society. vol. 1, pp. 21–26. Citeseer (2008)
8. Ding, S., Chen, J., Wang, Y., Kang, Y., Song, W., Cheng, J., Cao, Y.: E-mlb: Multilevel benchmark for event-based camera denoising. IEEE TMM pp. 1–12 (2023). https://doi.org/10.1109/TMM.2023.3260638
9. Duan, P., Wang, Z.W., Shi, B., Cossairt, O., Huang, T., Katsaggelos, A.K.: Guided event filtering: Synergy between intensity images and neuromorphic events for high performance imaging. IEEE TPAMI **44**(11), 8261–8275 (2022). https://doi.org/10.1109/TPAMI.2021.3113344
10. Duan, P., Wang, Z.W., Zhou, X., Ma, Y., Shi, B.: Eventzoom: Learning to denoise and super resolve neuromorphic events. In: CVPR. pp. 12819–12828 (2021). https://doi.org/10.1109/CVPR46437.2021.01263
11. Feng, Y., Lv, H., Liu, H., Zhang, Y., Xiao, Y., Han, C.: Event density based denoising method for dynamic vision sensor. Applied Sciences **10**(6), 2024 (2020), https://www.mdpi.com/2076-3417/10/6/2024
12. Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. IEEE TPAMI **44**(1), 154–180 (2022). https://doi.org/10.1109/TPAMI.2020.3008413, https://www.ncbi.nlm.nih.gov/pubmed/32750812
13. Graham, B., Engelcke, M., Maaten, L.v.d.: 3d semantic segmentation with submanifold sparse convolutional networks. In: CVPR. pp. 9224–9232 (2018). https://doi.org/10.1109/CVPR.2018.00961
14. Gu, D., Li, J., Zhu, L., Zhang, Y., Ren, J.S.: Reliable event generation with invertible conditional normalizing flow. IEEE TPAMI pp. 1–16 (2023). https://doi.org/10.1109/TPAMI.2023.3326538
15. Guo, S., Delbruck, T.: Low cost and latency event camera background activity denoising. IEEE TPAMI **45**(1), 785–795 (2023). https://doi.org/10.1109/TPAMI.2022.3152999

16. Guo, S., Kang, Z., Wang, L., Li, S., Xu, W.: Hashheat: An o(c) complexity hashing-based filter for dynamic vision sensor. In: 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC). pp. 452–457 (2020). https://doi.org/10.1109/ASP-DAC47756.2020.9045268
17. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1312–1321. https://doi.org/10.1109/CVPRW53098.2021.00144
18. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1312–1321 (2021). https://doi.org/10.1109/CVPRW53098.2021.00144
19. Khodamoradi, A., Kastner, R.: O(n)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors. IEEE Transactions on Emerging Topics in Computing **9**(1), 15–23 (2021). https://doi.org/10.1109/TETC.2017.2788865
20. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: A hierarchy of event-based time-surfaces for pattern recognition. IEEE TPAMI **39**(7), 1346–1359 (2017). https://doi.org/10.1109/TPAMI.2016.2574707
21. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits **43**(2), 566–576 (2008). https://doi.org/10.1109/JSSC.2007.914337, dVS128
22. Lin, S., Ma, Y., Guo, Z., Wen, B.: Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. pp. 578–593. Springer Nature Switzerland (2022)
23. Moeys, D.P., Corradi, F., Li, C., Bamford, S.A., Longinotti, L., Voigt, F.F., Berry, S., Taverni, G., Helmchen, F., Delbruck, T.: A sensitive dynamic and active pixel vision sensor for color or neural imaging applications. IEEE Transactions on Biomedical Circuits and Systems **12**(1), 123–136 (2018). https://doi.org/10.1109/TBCAS.2017.2759783, dAVIS346
24. Mohamed, S.A.S., Yasin, J.N., Haghbayan, M.H., Heikkonen, J., Tenhunen, H., Plosila, J.: Dba-filter: A dynamic background activity noise filtering algorithm for event cameras. In: Arai, K. (ed.) Intelligent Computing. pp. 685–696. Springer International Publishing (2022). https://doi.org/https://doi.org/10.1007/978-3-030-80119-9_44
25. Mueggler, E., Forster, C., Baumli, N., Gallego, G., Scaramuzza, D.: Lifetime estimation of events from dynamic vision sensors. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). pp. 4874–4881 (2015). https://doi.org/10.1109/ICRA.2015.7139876
26. Nozaki, Y., Delbruck, T.: Temperature and parasitic photocurrent effects in dynamic vision sensors. IEEE Transactions on Electron Devices **64**(8), 3239–3245 (2017). https://doi.org/10.1109/TED.2017.2717848, pixelVoltage
27. Nozaki, Y., Delbruck, T.: Temperature and parasitic photocurrent effects in dynamic vision sensors. IEEE Transactions on Electron Devices **64**(8), 3239–3245 (2017). https://doi.org/10.1109/TED.2017.2717848, pixelVoltage
28. Ojeda, F.C., Bisulco, A., Kepple, D., Isler, V., Lee, D.D.: On-device event filtering with binary neural networks for pedestrian detection using neuromorphic vision sensors. In: ICIP. pp. 3084–3088 (2020). https://doi.org/10.1109/ICIP40778.2020.9191148
29. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: NeurIPS. p. 5105–5114 (2017)

30. Ruedi, P.F., Heim, P., Kaess, F., Grenet, E., Heitger, F., Burgi, P.Y., Gyger, S., Nussbaum, P.: A 128 × 128 pixel 120-db dynamic-range vision-sensor chip for image contrast and orientation extraction. IEEE Journal of Solid-State Circuits **38**(12), 2325–2333 (2003). https://doi.org/10.1109/JSSC.2003.819169, dVS
31. Smith, J.: The Inverted Microscope: (a New Form of Microscope.) With the Description of a New Eye-piece Micrometer, and a New Form of Goniometer for Measuring the Angles of Crystals Under the Microscope (1852), https://books.google.com/books?id=UthvGwAACAAJ
32. Wang, Y., Du, B., Shen, Y., Wu, K., Zhao, G., Sun, J., Wen, H.: Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In: CVPR. pp. 6351–6360 (2019). https://doi.org/10.1109/CVPR.2019.00652
33. Wang, Z.W., Duan, P., Cossairt, O., Katsaggelos, A., Huang, T., Shi, B.: Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In: CVPR. pp. 1606–1616 (2020). https://doi.org/10.1109/CVPR42600.2020.00168
34. Wu, J., Ma, C., Yu, X., Shi, G.: Denoising of event-based sensors with spatial-temporal correlation. In: ICASSP. pp. 4437–4441 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053002
35. Zhang, P., Ge, Z., Song, L., Lam, E.Y.: Neuromorphic imaging with density-based spatiotemporal denoising. IEEE Transactions on Computational Imaging **9**, 530–541 (2023). https://doi.org/10.1109/TCI.2023.3281202
36. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in Neural Information Processing Systems. vol. 31 (2018)
37. Zhao, H., Jiang, L., Jia, J., Torr, P., Koltun, V.: Point transformer. In: ICCV. pp. 16239–16248 (2021). https://doi.org/10.1109/ICCV48922.2021.01595