


# S2N: Suppression-Strengthen Network for Event-based Recognition under Variant Illuminations

Zengyu Wan\*, Yang Wang\*, Ganchao Tan, Yang Cao, and Zheng-Jun Zha<sup>†</sup> 

University of Science and Technology of China, Hefei, China

{wanzengy, tgc1997}@mail.ustc.edu.cn

{ywang120, forrest, zhazj}@ustc.edu.cn

**Abstract.** The emerging event-based sensors have demonstrated outstanding potential in visual tasks thanks to their high speed and high dynamic range. However, the event degradation due to imaging under low illumination obscures the correlation between event signals and brings uncertainty into event representation. Targeting this issue, we present a novel suppression-strengthen network (S2N) to augment the event feature representation after suppressing the influence of degradation. Specifically, a suppression sub-network is devised to obtain intensity mapping between the degraded and denoised enhancement frames by unsupervised learning. To further restrain the degradation’s influence, a strengthen sub-network is presented to generate robust event representation by adaptively perceiving the local variations between the center and surrounding regions. After being trained on a single illumination condition, our S2N can be directly generalized to other illuminations to boost the recognition performance. Experimental results on three challenging recognition tasks demonstrate the superiority of our method. The codes and datasets could refer to <https://github.com/wanzengy/S2N-Suppression-Strengthen-Network>.

**Keywords:** Event camera, Event degradation, Event-based recognition, Motion evolution

## 1 Introduction

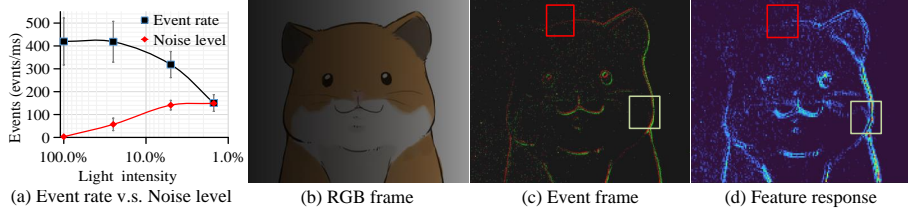
Event cameras are neuromorphic vision sensors that continuously encode the brightness changes of scene and asynchronously output a sequence of events as

$$I(x, y, t) - I(x, y, t - \Delta t) \geq p \cdot C, \quad (1)$$

where  $I(x, y, t)$  denotes the brightness on pixel location  $(x, y)$  at time  $t$ ,  $\Delta t$  is the time interval between adjacent event at the same pixel,  $p$  indicates the direction of the brightness change and  $C$  is the contrast sensitivity. Benefited from the difference imaging mechanism, event camera has many advantages (*e.g.*,

---

\* Equal contribution. <sup>†</sup> Corresponding author.



**Fig. 1.** (a) Measured event rate and noise level versus illumination intensity. The noise level is calculated by counting the number of events in a still scene. (b) The captured scene with traditional frame camera. (c) With the decrease of illumination intensity, the noise level of the event camera gradually increases and the real activity events gradually reduce, which will attenuate the amplitude of feature response and destroy the completeness of structure as in (d). The red box and yellow box in (c) and (d) marks area under different illumination conditions.

high dynamic range, low power consumption, and microsecond-scale temporal resolution [6, 9, 12, 23, 29, 49]), and shows great application potential in many visual and robotics tasks [2, 4, 7, 28, 30, 39, 44, 47, 48].

Despite the numerous advances of event-based vision, current event sensor prototypes, *e.g.*, DAVIS240, still bear unconventional sensor degradation [14, 19, 36]. As shown in Fig. 1 (a, c), the output of the event camera is susceptible to noise (*e.g.*, background activity noise), and the noise level is inversely proportional to illumination. In addition, the lower illumination also weakens the signal photocurrent and pixel bandwidth [16, 25, 31], resulting in the number reduction of activity events. The event degradation will obscure the correlation between event signals, resulting in errors and uncertainties in the generated representations, as shown in Fig. 1 (d).

Recently, some event-based recognition methods have been proposed to capture the spatial-temporal correlation from event stream with orderless aggregating or graph [5, 40, 45]. However, these methods ignore the influence of event degradation, which is inevitable in low light or low reflection environments. And this will lead to noisy and incomplete feature representations and result in recognition performance drop significantly. To address this problem, we propose a novel Suppression-Strengthen Network (S2N), which consists of two sub-networks: Noise Suppression Network (NSN) and Feature Strengthen Network (FSN), accounting for noise reduction and feature enhancement, respectively.

Specifically, in the first stage, the NSN is devised to obtain intensity mapping between the degraded frames and the denoised enhancement frames by unsupervised learning. To distinguish the real activity event and random noise, three novel discriminant loss functions are introduced by exploiting the fact that, in a small neighborhood, the events are originated from the identical motion, while noises are not. In the second stage, a novel FSN is presented to guarantee the completeness of the feature by motion evolution perception and local variations encoding. The Evolution Aware Module (EAM) from FSN is firstly de-

signed to perceive the motion evolution in each direction, resulting in a motion evolution map. Then through the Density-adaptive Central Differential Convolution(DCDC) process, the local center-surrounding variations of map are progressively encoded and adaptively aggregated into a complete event representation. Evaluations on three challenging event-based recognition tasks demonstrate the promising performance of our proposed method under variant illuminations. In summary, the main contributions of this paper are:

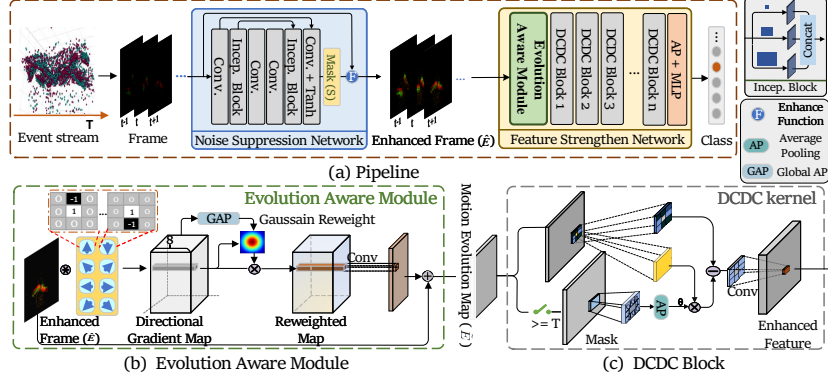
1. We propose the first S2N framework that considers event noise and real activity events reduction in event-based recognition to improve the inference accuracy and model robustness under various illumination conditions.
2. We devise a suppression sub-network to reduce random noise and enhance event frame contrast by learning an intensity mapping between the input and enhanced frames in an unsupervised manner. And subsequently, we propose a strengthen sub-network to generate robust event representation by perceiving the motion evolution with adaptive aggregation.
3. We collect two novel event-based datasets for gait and character recognition under variant illumination conditions, termed DAVIS346Gait and DAVIS346Character, respectively.
4. Our model can be well generalized to other illuminations after being trained in a single illumination scene, without fine-tuning. Extensive experiments on three challenging tasks show that our network outperforms the SOTA methods on various illumination conditions.

## 2 Related Work

In this section, we mainly review recent works, including event denoising and event-based recognition methods.

### 2.1 Event Denoising

One of the main challenges for event-based vision tasks is excessive noise. Considering the gap between events and RGB images [1, 12, 17, 34, 41], it is difficult to directly utilize the denoising method for RGB images in event-based work. Most existing event denoising methods work by exploiting the spatio-temporal correlation. Liu *et al.* [27] used the eight neighborhood pixels of an incoming event and considered it as noise if there lacks enough event support. Khodamoradi *et al.* [18] filtered BA (background activity) noise with compound memory cells, one for storing polarity and another for timestamp to improve the denoising process. Padala *et al.* [32] utilized the temporal correlation and spatial correlation and proposed filter with two layers to denoise. Wang *et al.* [42] tried to calculate the optical flow of events by local region fitting and remove the event above the threshold. Feng *et al.* [11] reduced the event noise based on the event density in the spatio-temporal neighborhood. The EDnCNN network [3] was developed for event denoising through labeling the probability of each event. These methods consider the noise effect but ignore the influence on activity event reduction, leading to an incomplete structure in the feature space.



**Fig. 2.** The overall structure of S2N. The event stream will be firstly converted into the even frames as input and be send to the noise suppression network for noise reduction and contrast enhancement in an unsupervised manner. Next, the enhanced frames are transmitted to the Evolution Aware Module to extract the motion evolution map. Then, the evolution map is put into the density-adaptive central differential convolution process to extract the complete event representation.

## 2.2 Event-based Recognition

Benefited from high-dynamic-range and high temporal resolution proprieties, the event camera is widely applied to assist recognition tasks. Lagorce *et al.* [10, 21, 22, 36] proposed event-based spatio-temporal features called time-surfaces and classified 36 characters by hand-crafted feature extractor. J.A.Pérez-Carrasco *et al.* [33] tried to accumulate the fixed-length event stream equally into an event frame and leveraged simple 4-layer CNN to learn to recognize the poker card symbol. Lee *et al.* [35, 37] proposed to use SNN (Spiking Neuron Network) to extract event stream features. However SNN lacks an effective way to train the network and needs to work on specific hardware. Wang *et al.* [40, 45] treated the event steam as point cloud and extracted the spatio-temporal structure with order-less aggregating through pointnet. Some recent works paid attention to graph convolution network [5, 24, 43] to extract spatio-temporal relations by building event graph. Also, a two-stage approach [42] was proposed to solve the noise interference by exploiting local motion consistency and desined a deep neural network to recognize gait from the denoised event stream. However, these recognition methods are susceptible to event degradation, while our work directs at enhancing the model’s robustness to event noise and real-activity event reduction.

## 3 Suppression-Strengthen Network

Our proposed Suppression Strengthen Network (S2N) contains two sub-networks: Noise Suppression Network (NSN) and Feature Strengthen Network (FSN),

which are responsible for noise reduction and feature enhancement, respectively as shown in Fig. 2. Given the event stream as input, we first accumulate the events within fixed time windows into two-channel frames according to the event polarity. The red and green colors of the event frames in Fig. 2 correspond to the event’s positive and negative polarity, respectively. As we have emphasized above, the event degradation includes event noise and activity events reduction, which will lead to excessive noise, lower contrast ratio, and structure break-points in event frames. Thus, we firstly send the event frames to NSN for noise removal and contrast enhancement in an unsupervised manner (Sec. 3.1). Next, the enhanced frames are transmitted to FSN to reconstruct the complete structure under the guidance of gradient prior and generate a motion evolution map, which is utilized to guide the density-adaptive central differential convolution for robust recognition (Sec. 3.2).

### 3.1 Noise Suppression Network

To achieve event denoising and contrast enhancement, we propose a novel unsupervised Noise Suppression Network (NSN), as shown in Fig. 2 (a). The NSN contains six layers mixed with plain convolution and inception block [38], and we adopt the Tanh function as the activation function to output a confidence mask (denoted as  $S$ ) of which the value belongs to  $-1$  to  $1$ , and the size is the same as input event frames. Under the constraints of the specifically designed loss functions, the weight of mask  $S$  will converge to  $-1$  value at the noise position for suppression, while to  $1$  at the activity events area for enhancement. With the confidence mask, the enhanced event frame can be obtained as

$$\hat{E}(x, y) = E(x, y) + S(x, y)E(x, y)(1 - E(x, y)), \quad (2)$$

where  $\hat{E}(x, y)$ ,  $E(x, y)$ ,  $S(x, y)$  represents the intensity of enhanced event frame, input event frame and the mask at coordinate  $(x, y)$ , respectively. Given  $E(x, y) \in (0, 1)$  and  $S(x, y) \in (-1, 1)$ . The Eq. (2) satisfy three constraints: 1. Capable of non-linear stretching to enhance the contrast; 2. Ensure the output is between  $[0, 1]$  without overflow truncation; 3. Monotonous and derivable. We choose the second-order formula  $E(1 - E)$  in Eq. (2) for easy implementation.

**Loss functions.** To ensure that the confidence mask can effectively suppress noise and enhance activity events, we propose three loss functions: noise suppression loss, enhancement loss, and consistency loss. The design of the loss functions is based on the observation: in a small neighborhood, the events are originated from the identical motion while the noises are not. Consequently, we can distinguish noise from input event signals by local spatial correlation.

**Noise Suppression Loss.** Based on the above observation, we use a  $3 \times 3$  kernel with the center weight of 0 and surrounding of 1 to filter the input event frame, and the output represents the strength of local correlation between the center pixel and surroundings. When the response amplitude exceeds a given threshold, the corresponding events are considered as real activity events; otherwise, as noise. In this paper, we set the threshold as the mean value of the whole

image intensity as in Eq. (3). After obtaining the noise distribution, we directly constrain the enhancement result  $\hat{E}(x)$  in the noise region to be 0:

$$\begin{aligned} \mathcal{L}_N &= \sum_{p \in \mathcal{P}} \hat{E}(p), \\ \mathcal{P} &= \{p \mid \sum_{i=1}^3 \sum_{j=1}^3 (\hat{E}_{i+p_x, j+p_y} W_{i,j}) \leq \text{Mean}(\hat{E})\}, \end{aligned} \quad (3)$$

where  $\mathcal{P}$  represents the set of noise location  $p$ ,  $W$  is a  $3 \times 3$  kernel with center weight as 0 and surrounding as 1.

**Enhancement Loss.** The event reduction under variant illumination conditions will result in low contrast of event frame. To this end, we propose enhancement loss to adaptively enhance the intensity of the event frame under different illumination conditions. The enhancement loss can be formulated as

$$\begin{aligned} \mathcal{L}_E &= \sum_{p \notin \mathcal{P}} (A - \sqrt{A})^2, \\ A &= \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \hat{E}(i, j), \end{aligned} \quad (4)$$

where  $k$  represents the size of the neighborhood and is set as 5 in this paper, and  $A$  represents the average intensity within the neighborhood. Based on Eq. (2), we can get that  $A \in (0, 1)$  and  $A < \sqrt{A}$ . Thus, under the constraint of Eq. (4), the intensity of event frames can gradually increase. Note that we only perform Eq. (4) on valid event positions and the value of  $\sqrt{A}$  is truncated to greater than 0.5 (empirical threshold) to avoid gradient explosion.

**Consistency Loss.** To keep the consistent of the enhancement of two polarities in event frames, we propose a consistent loss function, defined as following:

$$\mathcal{L}_C = \sum (Y^p - Y^n)^2, \quad (5)$$

where  $Y^p$  is the average intensity value of the positive polarity channel of mask  $S$ , and  $Y^n$  denotes the average intensity value of the negative channel.

**Total Loss.** The total loss function can be expressed as

$$\mathcal{L} = \mathcal{L}_N + \lambda_1 \mathcal{L}_E + \lambda_2 \mathcal{L}_C, \quad (6)$$

where  $\lambda_1$  and  $\lambda_2$  are trade-off weights. We will represent the effect of each loss item in the supplementary material.

### 3.2 Feature Strengthen Network

In this section, we propose the Feature Strengthen Network (FSN) as shown in Fig. 2. Particularly, an evolution guided density-adaptive central difference convolution scheme is proposed to progressively encode the local center-surrounding variation and adaptively aggregate the features into a complete event representation under the guidance of the motion evolution map.

**Motion Evolution Map.** Due to the influence of event degradation, the input event frames inevitably exist breakpoints, leading to incomplete feature representation. To repair the corrupted information, we exploit the global motion evolution feature to guide the feature enhancement. Specifically, we first calculate the 8 directional gradient projections of an event frame, as follows:

$$\nabla \hat{E}_i = W_i \cdot \hat{E}, \quad i = 1, 2, \dots, 8, \quad (7)$$

where  $W_i$  represents the  $i_{th}$   $3 \times 3$  directional kernel as shown in Fig. 2 (b). Then we perform Global Average Pooling (GAP) [26] on each gradient map to get the global evolution information as

$$Z_i = GAP(\nabla \hat{E}_i). \quad (8)$$

A larger pooling value corresponds to the direction of the main evolution pattern, while a smaller value corresponds to a direction that is the non-edge or broken edge region features which should be repaired. So to obtain the complete evolution pattern, we re-weight the directional gradient map by the Gaussian function and the normalized global information. The re-weighted map is then applied to the event frame to obtain the motion evolution map  $\tilde{E}$ :

$$\begin{aligned} \sigma_i &= softmax(Z_i) = \frac{e^{Z_i}}{\sum_{j=1}^8 e^{Z_j}}, \quad i = 1, \dots, 8, \\ \nabla \tilde{E}_i &= \nabla \hat{E}_i \exp(-(\|\nabla \hat{E}_i\| / \sigma_i)^2), \\ \tilde{E} &= \hat{E} + W_i \cdot \nabla \tilde{E}_i, \end{aligned} \quad (9)$$

where  $\sigma_i = softmax(Z_i)$  represents the normalized global information,  $W_i$  represents a  $1 \times 1$  convolution kernel and the Gaussian function helps to enhance the major evolution direction and suppress the weak evolution direction to diffuse the evolution tendency.

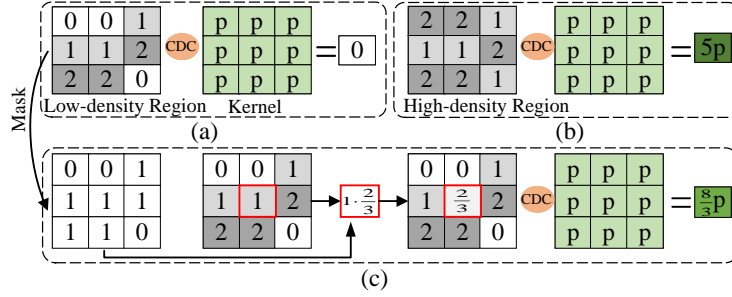
To guide subsequent operations, we need the evolution trend which is irrelevant of specific value. There, we perform a binary operation on the motion evolution map to obtain the trend mask  $M$ :

$$M(x, y) = \begin{cases} 1, & \tilde{E}(x, y) > T \\ 0, & \tilde{E}(x, y) \leq T, \end{cases} \quad (10)$$

where  $T$  represents the threshold and is set to 0 in this paper. Taking the motion evolution map  $\tilde{E}$  and mask  $M$  as input, we perform density-adaptive central difference convolution to extract the feature.

**Density-adaptive CDC kernel** Inspired by LBP(Local Binary Patterns), CDC [46] is devised to extract fine grained features by aggregating the center-oriented gradient of sampled values, which can be formulated as:

$$y_{ec}(p_0) = \sum_{p_n \in R} w(p_n) \cdot (x(p_n + p_0) - x(p_0)). \quad (11)$$



**Fig. 3.** (a) and (b) depict examples of CDC convolution acting in low-density regions and high-density regions, where density refers to the local proportion of valid values ( $> 0$ ), i.e., the average value of the evolution mask (also known as the motion evolution map binarization). The differential effect of CDC leads to excessive attenuation of low-density areas. (c) This problem can be suppressed by multiply the mean value of the mask with the central value to reduce the weight of the central difference term. To facilitate understanding, we adopt the simplest convolution core of the same  $p$ -value.

However, the difference term of CDC will lead to excessive reduction of the low-density region, as shown in Fig. 3. To solve this problem, we propose a density-adaptive central difference convolution. Specifically, we perform local average pooling on the evolution mask  $M$ , and get the local valid density  $\theta$ .

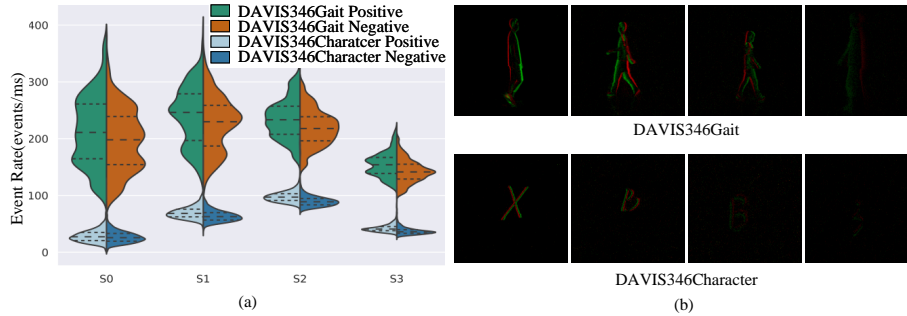
$$\theta(p_0) = \frac{1}{|R|} \sum_{p_n \in R} M(p_n + p_0), \quad (12)$$

where  $p_0$  is the central pixel,  $R$  is  $k \times k$  neighborhood of  $p_0$  and  $k$  is set to 3 there,  $|R|$  represents the area of  $R$ , the stride of the pooling is set 1 to keep the size unchanging. It is clear that pixels on structured regions tend to have larger density values, while the opposite is true for pixels on smooth regions. In each convolution process, we multiply  $\theta$  by the central pixel on the difference term, by adjusting the differential intensity to suppress the attenuation:

$$y_{ec}(p_0) = \sum_{p_n \in R} w(p_n) \cdot (x(p_n + p_0)) - \theta(p_0) \cdot x(p_0). \quad (13)$$

Besides, to maintain the feature response on smooth regions, we also introduce vanilla convolution and use the hyperparameter  $C \in [0, 1]$  to tradeoff the contribution between vanilla convolution and DCDC as

$$\begin{aligned} y(p_0) &= (1 - C) \cdot \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \\ &\quad + C \cdot \sum_{p_n \in R} w(p_n) \cdot (x(p_0 + p_n) - \theta(p_0)x(p_0)) \\ &= \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) - C \cdot \theta(p_0) \cdot x(p_0) \sum_{p_n \in R} w(p_n). \end{aligned} \quad (14)$$



**Fig. 4.** (a) The event rate of each instance in DAVIS346Gait and DAVIS346Character datasets. With the illumination intensity decrease, the output of activity events gradually decreases and noise level gradually increases. The inner lines of the violins represent the quarterlies. (b) The examples of DAVIS346Gait and DAVIS346Character datasets under four illumination conditions.

## 4 Experiment

In this section, we verify the effectiveness of each module in our network, evaluate the whole structure on three challenging event-based recognition tasks and analyze the performance.

### 4.1 Dataset

We adopt three datasets on three tasks to show the effectiveness of our method, which are all captured in the real scenes.

**DVS128Gesture.** The DVS128Gesture [2] is a gesture recognition benchmark captured by the DVS128 event camera, which contains 1,342 instances with 11 kinds of hand and arm gestures from 122 volunteers. DVS128Gesture is captured under five different illumination conditions, including fluorescent led, fluorescent, natural, led, and lab, and the illumination intensity is significantly different. For enough comparison, we segment the raw event stream through a 500ms sliding window and we get 13034 samples for training and 3546 for testing without the last class which corresponds to random gestures.

**DAVIS346Gait and DAVIS346Character.** To further evaluate the performance of our method on other recognition tasks, we perform experiments on gait recognition and character recognition tasks. However, the existing datasets for the two tasks are not captured under variant illuminations, thus not applicable for the evaluation. For example, the instances of DVS128Gait [43] are captured from different volunteers under two illumination conditions. To complement this, we present new gait (DAVIS346Gait) and character (DAVIS346Character) datasets captured under four kinds of illumination conditions, of which the light intensity is 300lux, 120lux, 15lux, and 6lux. Specifically, DAVIS346Gait contains 4,320 instances captured from 36 volunteers. And each volunteer is captured 30 times for each illumination condition. DAVIS346Character

**Table 1.** Ablation study results on DVSGesture dataset. S0: fluorescent led, S1: fluorescent, S2: natural, S3: led, S4: lab. We train each model on single illumination condition and test it on other four conditions. We report the mean accuracy on each condition. ‘w/o’ means without, ‘w/’ means with. (Acc %)

	Model	Train S4	Test			
			S0	S1	S2	S3
Loss Functions	NSN(w/o Le) + FSN	91.9	85.8(-11.4)	84.6(-11.5)	67.0(-25.1)	70.3(-21.2)
	NSN(w/o Ln) + FSN	96.0	95.9(-1.3)	93.3(-2.8)	86.6(-5.5)	88.0(-3.5)
	NSN(w/o Lc) + FSN	96.2	95.9(-1.3)	93.2(-2.9)	88.2(-3.9)	89.4(-2.1)
Modules	NSN + I3D	96.2	92.6(-4.6)	90.3(-5.8)	75.5(-16.6)	82.0(-9.5)
	FSN(w/ CDC)	96.4	90.6(-6.6)	89.9(-6.2)	74.5(-17.6)	78.9(-12.6)
	FSN(w/ DCDC)	96.9	91.2(-6.0)	91.6(-4.5)	77.4(-14.7)	79.6(-11.9)
	FSN(w/ EAM)	97.3	95.3(-1.9)	94.0(-2.1)	82.2(-9.9)	85.9(-5.6)
	FSN(w/ EAM & DCDC)	96.9	95.3(-1.9)	95.0(-1.1)	84.1(-8.0)	87.6(-3.9)
	NSN + FSN	<b>97.3</b>	<b>97.2</b>	<b>96.1</b>	<b>92.1</b>	<b>91.5</b>

**Table 2.** The results comparison on DVS128Gesture dataset. We train each model on single illumination condition and test it on all five conditions. We report the mean accuracy and variance after four conditions test. (Acc %)

Methods	S0 → all	S1 → all	S2 → all	S3 → all	S4 → all
I3D	<b>94.1±2.9</b>	92.7±5.2	91.4±2.3	93.0±2.3	84.5±9.1
ResNet34	82.6±7.8	83.9±8.1	85.7±5.5	85.3±2.6	72.2±13.1
PointNet++	86.7±2.3	87.6±3.4	85.0±2.7	87.2±2.5	83.3±5.8
EST	93.8±2.2	94.0±2.6	88.1±3.9	<b>93.3±5.1</b>	<b>85.3±8.0</b>
VN + I3D	93.9±1.5	<b>93.6±2.0</b>	<b>92.1±2.6</b>	92.1±1.9	83.0±8.3
IN + I3D	92.3±4.6	93.7±3.3	91.9±1.8	92.6±2.0	81.9±9.8
VN + EST	91.3±3.1	87.1±5.2	87.6±3.6	88.5±2.9	79.1±8.8
YN + EST	89.7±4.6	90.8±5.7	86.6±2.5	90.0±2.2	78.3±14.3
<b>S2N</b>	<b>97.2±0.8</b>	<b>94.6±1.4</b>	<b>95.3±2.2</b>	<b>94.6±1.4</b>	<b>94.7±2.8</b>

dataset is also captured under the same illumination condition, which contains 34,560 instances with 36 kinds of characters from 0-9 and A-Z. Due to space limitations, please refer to the supplemental material for more details about our data collecting process.

## 4.2 Implementation Details

The S2N is implemented in PyTorch and trained on an NVIDIA 3090 GPU. We use Adam [20] optimizer with learning rate periodically decaying from 1e-4 to 1e-6 and train each model for 25 epochs with a batch size of 16. The NSN will be pretrained firstly on the same training data and then the weight fixed. The  $\lambda_1$  and  $\lambda_2$  in Eq. (2) are set to 1 and 10 respectively to keep loss balance, and the C is 0.3. Note that FSN is adapted from different backbones in different tasks, specifically, I3D in gesture recognition, EV-Gait-IMG in gait recognition and ResNet34 in character recognition.

**Table 3.** The results on DAVIS346Gait dataset. L0, L1, L2 and L3 represent the illumination condition of 300lux, 120lux, 15lux and 6lux respectively. (Acc %)

Methods	L0 $\rightarrow$ all	L1 $\rightarrow$ all	L2 $\rightarrow$ all	L3 $\rightarrow$ all
EST	29.7	48.6	47.5	29.8
EV-Gait-3DGraph	25.9	29.2	21.7	19.6
VN + EV-Gait-IMG	68.6	56.8	<b>84.6</b>	51.6
YN + EV-Gait-IMG	<b>75.0</b>	<b>82.3</b>	75.0	<b>55.6</b>
VN + EST	51.8	49.1	47.3	26.5
YN + EST	39.5	63.6	58.2	26.2
<b>S2N</b>	<b>88.4</b>	<b>94.3</b>	<b>96.3</b>	<b>90.3</b>

**Table 4.** The results on DAVIS346Character dataset. L0, L1, L2 and L3 represent the illumination condition of 300lux, 120lux, 15lux and 6lux respectively. (Acc %)

Methods	L0 $\rightarrow$ all	L1 $\rightarrow$ all	L2 $\rightarrow$ all	L3 $\rightarrow$ all
ResNet34	27.0	70.0	<b>90.5</b>	78.5
EST	29.4	<b>90.7</b>	90.0	<b>88.7</b>
VN + ResNet34	34.8	75.3	76.4	68.6
YN + ResNet34	30.9	66.7	63.3	52.1
VN + EST	<b>47.7</b>	83.2	86.0	77.2
YN + EST	34.7	51.4	58.9	44.1
<b>S2N</b>	<b>78.9</b>	<b>91.4</b>	<b>95.7</b>	<b>92.1</b>

### 4.3 Ablation Study

We use the instances of *lab scene* in DVSGesture as training set and the other four scenes as test sets. Table 1 shows the results from different combinations of loss functions in NSN training and modules in our network. The result shows that each loss term contributes to better recognition results, with the exposure loss being the most important. And the EAM(Evolution-aware Module) and DCDC(Density-adaptive CDC) can strengthen feature effectively and consistently improve performance. Combining with the NSN(Noise Suppression Network), our whole network performs best, which again validates the effectiveness of each proposed module under variant illumination scenes.

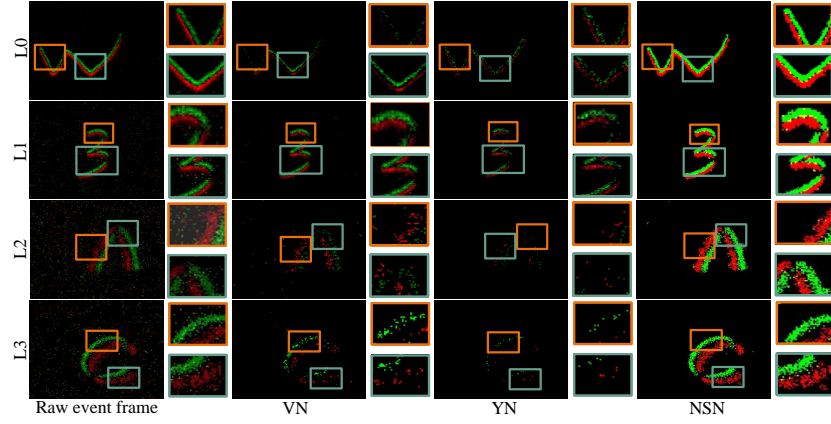
### 4.4 Comparisons Against SOTA methods

We compare our S2N against SOTA methods on three challenging event-based recognition tasks.

**Performance on DVSGesture.** We compare our method with four SOTA recognition methods: I3D [8], ResNet34 [15], PointNet++ [40], and EST [13]. Besides, to demonstrate the effectiveness of our NSN, we also compare our S2N with SOTA denoising methods: VN [42], and YN [11] and report the results of the denoising methods combined with recognition models, including I3D and EST. We train each model on a single illumination condition and examine the performance on the other four illuminations, respectively. The results are shown in

**Table 5.** Denoising methods comparison. To be fair, we uniformly use I3D as the feature extractor to compare the impact of the denoising algorithm on feature extraction. We train each model on a single illumination condition and test it on the other four illuminations. S0: fluorescent led, S1: fluorescent, S2: natural, S3: led, S4: lab. (Acc %)

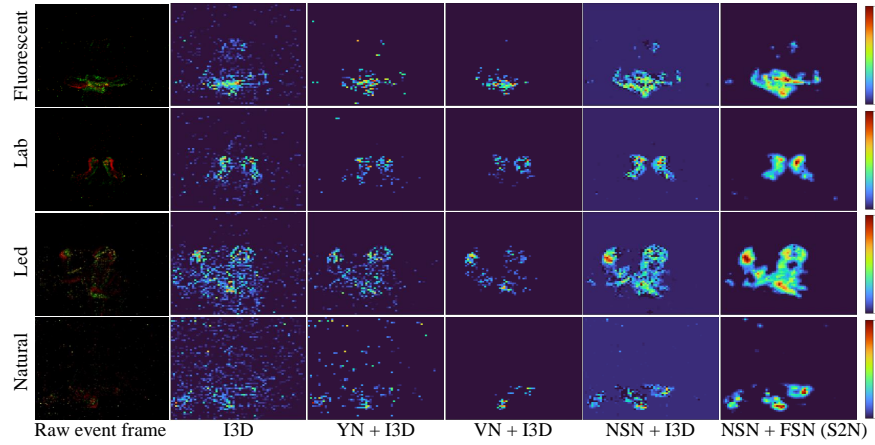
Methods	Train	Test			
	S4	S0	S1	S2	S3
I3D	97.5	90.3	87.0	75.0	78.4
VN + I3D	94.2	86.5(-3.8)	86.8(-0.2)	74.1(-0.9)	76.2(-2.2)
YN + I3D	97.3	83.7(-6.6)	84.1(-2.9)	70.5(-4.5)	77.9(-0.5)
NSN + I3D	96.2	92.6(+2.3)	90.3(+3.3)	75.5(+0.5)	82.0(+3.6)



**Fig. 5.** Denoising comparison of different methods on four illumination conditions. L0, L1, L2 and L3 represent the illumination condition of 300lux, 120lux, 15lux and 6lux respectively. Compared with VN and YN, our NSN effectively removes the noise while preserving the original activity events.

Table 2. Our method is 1.9% higher than the SOTA method I3D in recognition accuracy. It demonstrates that our proposed S2N can effectively suppress the influence of noise and strengthen the robustness of feature representation. Besides, compared with the combination of event denoising and recognition methods, our improvement is at least 1.3 % which is also significant. Furthermore, the overall variance of our results is smaller than other methods, which demonstrates the stability of our method.

**Performance on DAVIS346Gait.** We compare our method with SOTA gait recognition methods: EV-Gait-3DGraph and EST, as well as EV-Gait-IMG and EST combined with VN and YN. We randomly sample 50% data in each class and use it as the training set and the rest data as the test set. We train our model on one illumination condition and test it on other illuminations. The results are shown in Table 3. As can be seen, our method achieves the best performance under all illumination conditions.



**Fig. 6.** The feature map visualization of different methods including I3D, YN + I3D, VN + I3D, NSN + I3D and our model S2N. The feature maps are from the output of the first layer of models. We select the features of the same channel, normalize, and then display them in pseudo color for visualization.

**Performance on DAVIS346Character.** We compare our method with two character recognition methods: ResNet34 [15] and EST [13], as well as ResNet34 and EST combined with VN and YN. Like in DVSGesture and DAVIS-346Gait, We train each model on one illumination condition and examine the performance on the other four illuminations, respectively. The results shown in Table 4 validate the excellent performance of our method.

The above results demonstrate that our proposed method can be generalized to various recognition tasks. Furthermore, our method can be directly used to improve the performance under variant illumination conditions without fine-tuning. While the performance of point-distribution-dependent methods, like EV-gait-3DGraph and Pointnet++, drop significantly because of the event distribution difference between variant illuminations.

#### 4.5 Performance Analysis

**Noise suppression results** To demonstrate the effectiveness of our proposed NSN, we compare our NSN with VN and YN qualitatively and quantitatively. From Fig. 5, we can observe that although VN can remove the noise to a certain extent, the enhanced results still have noise residual. While the YN algorithm can efficiently remove the unfavorable noise, the activity events are also obscured. Different from VN and YN, our NSN effectively removes the noise while keeping the original activity events unchanged. Furthermore, our NSN can also enhance the intensity, which boosts feature response and improve the recognition accuracy, as shown in Table 5. As the illumination scene changes, our NSN can stably remove the noise and enhance frame intensity to help to extract features, which shows our method’s generalization ability under variant illuminations.

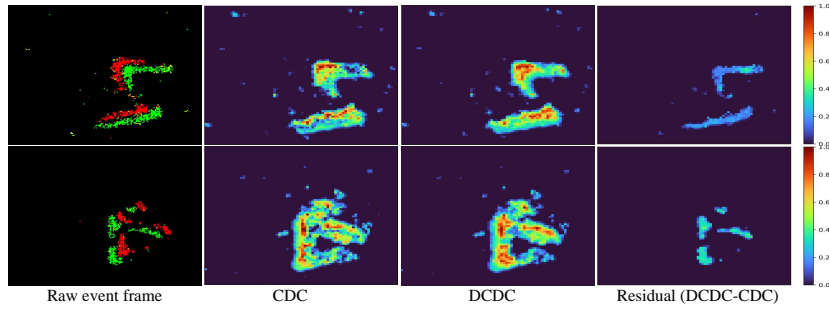


Fig. 7. The guiding role of motion evolution map.

**Feature strengthen results** We show the degraded feature and the feature enhancement results of different methods on four illumination conditions of DVS-Gesture in Fig. 6. Due to the influence of noise, the amplitude of feature response on discriminative regions is weakened and full of noise. The event denoising algorithm can suppress the influence of noise. However, these algorithms cannot significantly improve the amplitude of feature response. Compared with them, our method significantly improves the feature response amplitude on discriminative regions while suppressing the noise interference. Furthermore, we compare the feature representation between DCDC and CDC as in Fig. 7, which shows that the structural information will be well restored under the guidance of the motion evolution map.

## 5 Conclusion

In this paper, we propose a novel suppression-strengthen network for event-based recognition under variant illumination conditions. The NSN sub-network is first proposed to effectively reduce random noise and enhance the contrast of the event frame in an unsupervised manner. And then, the FSN sub-network progressively encodes the motion evolution information and aggregates the information into a complete representation. After being trained on one single illumination condition, our S2N can be generalized to other illumination conditions. Extensive experiments on different illumination conditions and three recognition tasks validate the effectiveness of our proposed method.

## 6 Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, National Natural Science Foundation of China (NSFC) under Grants U19B2038 and 61872327, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025.

## References

1. 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) [3](#)
2. Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., Nayak, T., Andreopoulos, A., Garreau, G., Mendoza, M., Kusnitz, J., Debole, M., Esser, S., Delbruck, T., Flickner, M., Modha, D.: A Low Power, Fully Event-Based Gesture Recognition System. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI (2017) [2](#), [9](#)
3. Baldwin, R., Almatrafi, M., Asari, V., Hirakawa, K.: 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. pp. 1701–1710 (2020) [3](#)
4. Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous Optical Flow and Intensity Estimation from an Event Camera. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) [2](#)
5. Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., Andreopoulos, Y.: Graph-Based Object Classification for Neuromorphic Vision Sensing. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, Korea (South) (2019) [4](#)
6. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A  $240 \times 180$  130 db  $3 \mu\text{s}$  latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits* **49** (2014) [2](#)
7. Calabrese, E., Taverni, G., Easthope, C.A., Skriabine, S., Corradi, F., Longinotti, L., Eng, K., Delbruck, T.: DHP19: Dynamic Vision Sensor 3D Human Pose Dataset. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, Long Beach, CA, USA (2019) [2](#)
8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) [11](#)
9. Chen, J., Wang, Y., Cao, Y., Wu, F., Zha, Z.J.: Progressivemotionseg: Mutually reinforced framework for event-based motion segmentation. *arXiv preprint arXiv:2203.11732* (2022) [2](#)
10. Clady, X., Maro, J.M., Barré, S., Benosman, R.B.: A Motion-Based Feature for Event-Based Pattern Recognition. *Frontiers in Neuroscience* **10** (2017) [4](#)
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** (2020) [3](#), [11](#)
12. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* **44**(1), 154–180 (2020) [2](#), [3](#)
13. Gehrig, D., Loquercio, A., Derpanis, K., Scaramuzza, D.: End-to-End Learning of Representations for Asynchronous Event-Based Data. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, Korea (South) (2019) [11](#), [13](#)
14. Guo, S., Kang, Z., Wang, L., Zhang, L., Chen, X., Li, S., Xu, W.: A noise filter for dynamic vision sensors using self-adjusting threshold. *arXiv preprint arXiv:2004.04079* (2020) [2](#)

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs] (2015) [11](#), [13](#)
16. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1312–1321 (2021) [2](#)
17. Huang, Y., Zha, Z.J., Fu, X., Zhang, W.: Illumination-invariant person re-identification. In: Proceedings of the 27th ACM international conference on multimedia. pp. 365–373 (2019) [3](#)
18. Khodamoradi, A., Kastner, R.: O(N)-Space Spatiotemporal Filter for Reducing Noise in Neuromorphic Vision Sensors. IEEE Transactions on Emerging Topics in Computing **9** (2021) [3](#)
19. Kim, J., Bae, J., Park, G., Zhang, D., Kim, Y.M.: N-imagenet: Towards robust, fine-grained object recognition with event cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2146–2156 (October 2021) [2](#)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) [10](#)
21. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: HOTS: A Hierarchy of Event-Based Time-Surfaces for Pattern Recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017) [4](#)
22. Lee, J.H., Delbruck, T., Pfeiffer, M., Park, P.K.J., Shin, C.W., Ryu, H., Kang, B.C.: Real-Time Gesture Interface Based on Event-Driven Processing From Stereo Silicon Retinas. IEEE Transactions on Neural Networks and Learning Systems **25** (2014) [4](#)
23. Li, C., Brandli, C., Berner, R., Liu, H., Yang, M., Liu, S.C., Delbruck, T.: Design of an RGBW color VGA rolling and global shutter dynamic and active-pixel vision sensor. In: 2015 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, Lisbon, Portugal (2015) [2](#)
24. Li, Y., Zhou, H., Yang, B., Zhang, Y., Cui, Z., Bao, H., Zhang, G.: Graph-based asynchronous event processing for rapid object recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 934–943 (October 2021) [4](#)
25. Lichtsteiner, P., Posch, C., Delbruck, T.: A  $128 \times 128$  120 dB 15  $\mu$ s latency asynchronous temporal contrast vision sensor. IEEE journal of solid-state circuits **43** (2008) [2](#)
26. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013) [7](#)
27. Liu, H., Brandli, C., Li, C., Liu, S.C., Delbruck, T.: Design of a spatiotemporal correlation filter for event-based sensors. In: 2015 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE (2015) [3](#)
28. Maqueda, A.I., Loquercio, A., Gallego, G., Garcia, N., Scaramuzza, D.: Event-Based Vision Meets Deep Learning on Steering Prediction for Self-Driving Cars. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT (2018) [2](#)
29. Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–9. IEEE (2018) [2](#)
30. Moeys, D.P., Corradi, F., Kerr, E., Vance, P.J., Das, G.P., Neil, D., Kerr, D., Delbrück, T.: Steering a predator robot using a mixed frame/event-driven convolutional neural network. 2016 Second International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP) (2016) [2](#)

31. Nozaki, Y., Delbruck, T.: Temperature and Parasitic Photocurrent Effects in Dynamic Vision Sensors. *IEEE Transactions on Electron Devices* **64** (2017) 2
32. Padala, V., Basu, A., Orchard, G.: A noise filtering algorithm for event-based asynchronous change detection image sensors on TrueNorth and its implementation on TrueNorth. *Frontiers in neuroscience* **12** (2018) 3
33. Perez-Carrasco, J.A., Bo Zhao, Serrano, C., Acha, B., Serrano-Gotarredona, T., Shouchun Chen, Linares-Barranco, B.: Mapping from Frame-Driven to Frame-Free Event-Driven Vision Systems by Low-Rate Rate Coding and Coincidence Processing—Application to Feedforward ConvNets. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** (2013) 4
34. Ramesh, B., Yang, H., Orchard, G., Le Thi, N.A., Zhang, S., Xiang, C.: DART: Distribution Aware Retinal Transform for Event-Based Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42** (2020) 3
35. Shrestha, S.B., Orchard, G.: SLAYER: Spike Layer Error Reassignment in Time. In: *Advances in Neural Information Processing Systems*. vol. 31. Curran Associates, Inc. (2018) 4
36. Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., Benosman, R.: Hats: Histograms of averaged time surfaces for robust event-based object classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1731–1740 (2018) 2, 4
37. Stromatias, E., Neil, D., Galluppi, F., Pfeiffer, M., Liu, S.C., Furber, S.: Scalable energy-efficient, low-latency implementations of trained spiking Deep Belief Networks on SpiNNaker. In: *2015 International Joint Conference on Neural Networks (IJCNN)* (2015) 4
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2818–2826 (2016) 5
39. Tan, G., Wang, Y., Han, H., Cao, Y., Wu, F., Zha, Z.J.: Multi-grained spatio-temporal features perceived network for event-based lip-reading. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 20094–20103 (June 2022) 2
40. Wang, Q., Zhang, Y., Yuan, J., Lu, Y.: Space-Time Event Clouds for Gesture Recognition: From RGB Cameras to Event Cameras. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Waikoloa Village, HI, USA (2019) 2, 4, 11
41. Wang, Y., Cao, Y., Zha, Z.J., Zhang, J., Xiong, Z., Zhang, W., Wu, F.: Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement. In: *Proceedings of the 27th ACM international conference on multimedia*. pp. 2015–2023 (2019) 3
42. 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: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA (2019) 3, 4, 11
43. Wang, Y., Zhang, X., Shen, Y., Du, B., Zhao, G., Cui Lizhen, L.C., Wen, H.: Event-Stream Representation for Human Gaits Identification Using Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) 4, 9
44. Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 2563–2572 (October 2021) 2

45. Yang, J., Zhang, Q., Ni, B., Li, L., Liu, J., Zhou, M., Tian, Q.: Modeling point clouds with self-attention and gumbel subset sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3323–3332 (2019) [2](#), [4](#)
46. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5295–5305 (2020) [7](#)
47. Zhang, J., Yang, X., Fu, Y., Wei, X., Yin, B., Dong, B.: Object tracking by jointly exploiting frame and event domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13043–13052 (2021) [2](#)
48. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Long Beach, CA, USA (2019) [2](#)
49. Zou, S., Guo, C., Zuo, X., Wang, S., Wang, P., Hu, X., Chen, S., Gong, M., Cheng, L.: Eventhpe: Event-based 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10996–11005 (October 2021) [2](#)