# Temporal-Guided Spiking Neural Networks for Event-Based Action Recognition

BMVC 2023 Submission # 137

**Abstract**

This paper explores the promising interplay between spiking neural networks (SNNs) and event-based cameras for privacy-preserving human action recognition (HAR). The innate ability of event cameras to capture exclusively the contours of moving subjects, coupled with the inherent capability of SNNs in transmitting spatiotemporal information through discrete spikes, establishes a highly synergistic compatibility between these two technologies for achieving event-based HAR. However, previous studies on event-based HAR have relied solely on spiking neurons to handle long-term temporal information, which is crucial for accurate HAR, limiting SNNs' performance in this area. In this paper, we introduce two novel frameworks aimed at augmenting the ability of SNNs to process long-term temporal information: temporal segment-based SNN (*TS-SNN*) and 3D convolutional SNN (*3D-SNN*). The *TS-SNN* extracts long-term temporal information by dividing actions into shorter segments, while the *3D-SNN* replaces 2D spatial elements with 3D components to facilitate the transmission of temporal information. To promote further research in event-based HAR, we create a dataset, *FallingDetection-CeleX*, collected using the high-resolution CeleX-V event camera $(1280 \times 800)$, comprising 7 distinct actions. Extensive experimental results show that our proposed frameworks surpass state-of-the-art SNN methods on our newly collected dataset and three other neuromorphic datasets, showcasing their effectiveness in handling long-range temporal information for event-based HAR.

# 1 Introduction

Spiking neural networks (SNNs) represent the third generation [11, 26] of neural networks and are distinguished for their ability to perform tasks with ultra-low power consumption when deployed on dedicated neuromorphic hardware [32, 47]. These networks transmit spatiotemporal information between units via discrete spikes, mimicking the biological neural system. SNNs are inherently compatible with event-based cameras, which employ bio-inspired sensors to asynchronously measure per-pixel brightness fluctuations, thereby generating an event stream that encodes time, location, and sign of the brightness changes [9]. Consequently, SNNs are ideally suited for integration with event cameras.

Previous research on SNNs [22, 31, 44] has showcased impressive accomplishments in object recognition and classification tasks. We posit that the integration of SNNs and event-based cameras holds exceptional promise for human action recognition (HAR) [38]. Traditional video-based HAR models often raise privacy concerns [5, 6, 43], rendering their application inappropriate in private settings such as fall detection in bathrooms. SNNs working in tandem with event cameras can overcome this limitation as event cameras only capture the outline of moving subjects, ignoring their identifying features and static backgrounds, thereby preserving privacy while performing HAR.

Several studies [7, 8, 23] have demonstrated the applicability of SNNs for HAR on neuromorphic datasets captured by event cameras. However, their capability to manage long-range temporal information is solely dependent on spiking neurons within the SNNs, which is insufficient for video-based HAR. Effective processing of long-range temporal information is critical for accurate video-based HAR. In light of this, we propose two frameworks to improve the SNNs' capacity to process long-term temporal information. The first framework, temporal segment-based SNN (*TS-SNN*), implements a temporal segment strategy [39] on SNNs, which enables the extraction of long-term temporal information by breaking down lengthy actions into shorter moments. The second framework, referred to 3D convolutional SNN (*3D-SNN*), involves substituting the 2D spatial components in SNNs with 3D spatial-temporal components to facilitate the transmission of temporal information between layers.

Furthermore, we collect a event-based HAR dataset, *FallingDetection-CeleX*, specifically focused on privacy-preserving applications to encourage further research in this area. At present, there is a scarcity of event-based real-world HAR datasets and existing datasets [1, 23, 28] primarily focus on standard action recognition scenarios, neglecting the most common privacy-preserving situation occurring in home settings, such as fall detection. Additionally, the majority of these event-based HAR datasets were captured using DVS128 and DAVIS346 sensors [20], which have relatively low resolutions of $128 \times 128$ and $346 \times 260$ pixels, respectively. In contrast, we utilize the CeleX-V [3] event camera for our recordings. The CeleX-V is a 1-megapixel multifunctional sensor with a high resolution of $1280 \times 800$ pixels, enabling it to capture more detailed information compared to lower-resolution alternatives. Our dataset consists of 875 recordings featuring 51 subjects performing 7 distinct actions, including three types of falls.

We conduct quantitative comparisons with state-of-the-art (SOTA) SNN methods on our newly collected dataset, as well as three additional standard neuromorphic datasets. The experimental results indicate that the proposed frameworks outperform SOTA methods, showing the effectiveness in processing long-range temporal information for event-based HAR.

The contributions of this work are fourfold. *First,* We introduce *TS-SNN*, which leverages temporal segment strategies to extract long-term temporal information. This method involves breaking down lengthy actions into shorter segments, enabling SNNs to process and analyze long-term temporal information more efficiently. *Second,* We propose *3D-SNN*, which substitutes the 2D spatial components in SNNs with 3D spatial-temporal components. By incorporating this change, the transmission of temporal information in SNNs is facilitated, leading to better processing of long-term temporal information for event-based HAR. *Third,* We collect an event-based action recognition dataset named *FallingDetection-Celex*, with a special focus on falling detection. The dataset includes 875 recordings of 51 subjects performing 7 distinct actions, including three categories of various types of falling. This dataset will be made publicly accessible to the community. *Last,* We evaluate two proposed frameworks using our *FallingDetection-CeleX* dataset as well as three additional challenging event-based HAR datasets. The results show that both frameworks outperform the existing SOTA accuracies across all four datasets.

# 2 Related Work

## 2.1 Event-based Action Recognition

Human action recognition (HAR) has attracted substantial interest from the academic community due to its various real-world applications, such as human-robot interaction [35], visual surveillance systems [24, 34], elderly person monitors systems [2, 21], and autonomous navigation systems [25]. Recently, event cameras have emerged as bio-inspired sensors that capture movements in the environment without compromising sensitive information. Therefore, event sensors are the perfect choice for privacy-preserving HAR. Innocenti *et al*. [15] converted the output of an event camera into frames and used standard computer vision techniques to analyze them. However, this method primarily concentrates on aggregating events and handling frames.

On the other hand, some researchers have proposed dealing directly with events. For instance, Maro *et al*. [27] introduced a framework for dynamic gesture recognition based on time surfaces developed by [18]. SNNs are ideal for processing event data as they transmit information through discrete spikes. George *et al*. [10] presented an SNN that utilizes convolution and reservoir computing to classify human hand gestures. Liu *et al*. [23] proposed a hierarchical SNN architecture for event-based action recognition, which leverages motion information. Additionally, Fang *et al*. [7] introduced the spike-element-wise (SEW) ResNet by applying residual learning to deep SNNs.

However, these models rely solely on spiking neurons to manage long-term temporal information. Nevertheless, their performance is not optimal. To address this issue, we propose two frameworks that incorporate spiking neurons to further enhance the ability of SNNs to process long-range temporal information.

## 2.2 Event-based Datasets

The availability of real-world event-based action recognition datasets is currently limited. Amir et al. [1] proposed a dataset for event-based hand gesture recognition, which was recorded using a dynamic vision sensor (DVS) [20]. Maro et al. [27] collected datasets for event-based human gestures utilizing an asynchronous time-based image sensor (ATIS) [33]. Additionally, Miao et al. [28] produced a neuromorphic dataset with a DAVIS camera from three different perspectives. However, the dataset is relatively small, including only 291 recordings.

More recently, Liu et al. [23] proposed a new dataset for event-based action recognition using a DVS camera with two lighting conditions and two camera positions. It is worth noting that the existing datasets were recorded using DVS128 and DAVIS346 sensors with resolutions of $128 \times 128$ and $346 \times 260$ pixels, respectively, which are relatively low resolutions. To ensure a more comprehensive and detailed capture of information, we use the CeleX-V [3] event camera instead for recording. The CeleX-V has a significantly higher resolution of $1280 \times 800$ pixels.

Moreover, all the existing datasets only focus on normal action recognition scenarios and overlook the most common privacy-preserving applications of action recognition, such as elder falling detection in homes. Hence, we gather a new event-based falling detection dataset. The newly acquired dataset is called *FallingDetection-CeleX*, which contains 875 recordings of 51 subjects performing seven different actions, including three different types of falling. The dataset is intended for both event-based action recognition and falling detection.

# 3 Preliminary: Spiking neuron model

Spiking neuron models serve as the fundamental computational unit of SNNs. A unified model [8] characterizes the dynamics of diverse spiking neuron types, employing the subsequent discrete-time equations:

$$H(t) = f(V(t-1), X(t)), \tag{1}$$

$$S(t) = \Theta(H(t) - V_{\text{th}}), \tag{2}$$

$$V(t) = H(t) \cdot (1 - S(t)) + V_{\text{reset}} \cdot S(t), \tag{3}$$

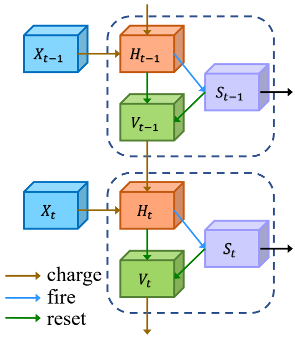


Figure 1: The general discrete spiking neuron model.

where $V(t)$ indicates the membrane potential after the trigger of a spike at time $t$, $H(t)$ represents the membrane potential after neuronal dynamics, $X(t)$ denotes the external input to the neuron at time $t$, $S(t)$ denotes the output spike at time $t$, and $\Theta(\cdot)$ is the Heaviside step function. Once the membrane potential $H(t)$ of a neuron reaches a particular threshold $V_{\text{th}}$ at a given time $t$, the neuron will generate a spike, resulting in the membrane potential dropping to a reset value $V_{\text{reset}}$ that is lower than the threshold $V_{\text{th}}$. This process is referred to as the hard reset and is widely used in deep SNNs. These equations constitute a general discrete spiking neuron model, which is illustrated in Figure 1.

Spiking neuron models that are commonly utilized include the Hodgkin-Huxley [14], Izhikevich [16], and leaky integrate-and-fire (LIF) [12] models. Apart from the variability in neuronal dynamics (Eq. 1) for distinct spiking neurons, all spiking neurons exhibit identical neuronal fire (Eq. 2) and reset (Eq. 3) equations. Among these models, the LIF model is the most straightforward and efficient, making it the ideal choice for implementation. The neuronal dynamics of the LIF neuron model are defined by [12]:

$$H(t) = V(t-1) + \frac{1}{\tau} \cdot (X(t) - (V(t-1) - V_{\text{reset}})), \tag{4}$$

where $\tau$ represents the membrane time constant. Fang *et al.* [8] have introduced a training algorithm that enables the learning of both the synaptic weights and membrane time constants of LIF spiking neurons, known as parametric leaky integrate-and-fire (PLIF) neurons. In this paper, we leverage the capability of PLIF neurons as the computational units to enhance the overall expressiveness of SNNs. The neuronal dynamics is defined by:

$$H(t) = V(t-1) + \frac{1}{1+\exp(-a)} (X(t) - (V(t) - V_{\text{reset}})), \tag{5}$$

where $a$ is a learnable parameter.

# 4 Method

Although SNNs have the ability to utilize temporal information, they still face challenges when processing long videos. In order to enhance their effectiveness in recognizing actions within videos, we have drawn inspiration from traditional video processing techniques and put forth two frameworks: *TS-SNN* (Section 4.2) and *3D-SNN* (Section 4.3). These frameworks aim to enhance SNNs' capacity to extract long-range temporal information, thereby improving their ability in handling lengthier videos.

## 4.1 Neuromorphic Preprocessing

Our research primarily centers around neuromorphic action recognition datasets directly captured from event cameras, which are typically processed using event-to-frame integration techniques [4, 13, 17, 19, 29, 30, 41, 42, 46] that convert event data into two-channel videos. The event data, represented as $e(x_i, y_i, t_i, p_i)(i = 0, 1, \cdots, N-1)$, captures the pixel location of brightness changes ($x_i$ and $y_i$), the timestamp ($t_i$), and the polarity ($p_i$) of the event. This data is then split into $T$ slices with a similar number of events in each slice before being integrated into frames. The final tensor produced has the shape of $[T, 2, H, W]$, where $H$ and $W$ indicate the height and width of the frames, respectively.

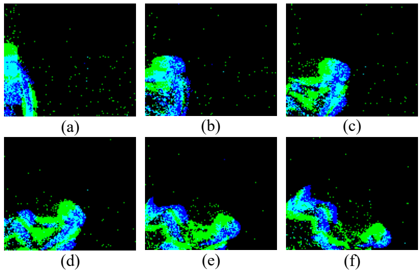


Figure 2: Video frames converted from event data, depicting a person falling down as captured from a side view. The frames are arranged in sequential order, from (a) to (f).

Subsequently, the event data is converted into videos with only two channels, and the number of frames is a hyperparameter that can be adjusted. These videos solely capture dynamic actions and omit the static background and detailed features of individuals, as depicted in Figure 2. This approach effectively preserves the privacy of users while still providing essential information about their actions.

## 4.2 Temporal Segment-Based Spiking Neural Network

Long-range temporal information is crucial in achieving high accuracy in human action recognition. A major limitation of conventional SNNs is that other components besides spiking neurons are unable to model long-term temporal information effectively. This is primarily due to their limited access to temporal context since they are designed to operate solely on a single frame, which is characteristic of spatial networks. However, complex actions, such as falling down, comprise multiple stages spanning over a relatively long period, and failing to utilize long-term temporal structures in SNNs' training would be a significant loss.

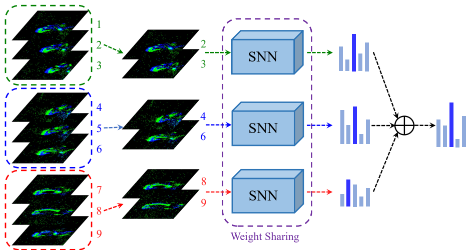


Figure 3: The framework of *TS-SNN*. As an example, a video comprising $N = 9$ frames is partitioned into $L = 3$ segments, and $K = 2$ frames are randomly selected from each segment. These segments are then processed through a weight-shared SEW ResNet, and the resulting distributions are combined to make predictions.

Therefore, as depicted in Figure 3, we propose the incorporation of the successful strategy of temporal segment network (TSN), traditionally used in video action recognition, into SNNs. Since the spatial components in SNNs are inherently designed to operate on a single frame, their prediction of short-term temporal information is more precise than long-term temporal information. Consequently, we divide the preprocessed video into $L$ segments, denoted by $\{S_1, S_2, ..., S_L\}$, with the aim of segmenting the long-range temporal information into shorter segments. From each segment, we randomly select $K$ frames belonging to the same moment of one action, which is solely processed by one SNN to obtain a more accurate prediction of the short-term temporal information. The SNNs processing different segments share weights, thus resulting in $L$ accurate short-term prediction distributions. Finally, these short-term distributions are combined using a straightforward fusion method such as summation, averaging, or maximum operation to obtain the ultimate accurate long-term distribution. This approach enhances the ability of SNNs to handle long-term temporal information, thereby improving the accuracy of event-based human action recognition.

Formally, the ultimate accurate long-term distribution is formulated as follows:

$$y = \mathrm{Softmax}\left(\upsilon(\varphi(S_1;\mathbf{W}), \varphi(S_2;\mathbf{W}), \cdots, \varphi(S_L;\mathbf{W}))\right), \tag{6}$$

where $S_i$ indicates the $i$th segment comprising $K$ frames. $\varphi(S_i;\mathbf{W})$ is the weight-shared SNN that processes each segment $S_i$ and produces its corresponding prediction distribution. The segmental consensus function $\upsilon$ combines the distributions from multiple segments to obtain a consensus of class prediction $y$ among them. This approach enhances the ability of SNNs to handle long-term temporal information, thereby improving the accuracy of event-based human action recognition. We refer to this proposed temporal segment-based method as *TS-SNN*.

## 4.3 3D Spiking Neural Network

Although the temporal segment strategy has significantly improved the performance of SNNs in event-based human action recognition, the spatial components of the SNN architecture still process frames one by one, resulting in temporal information delivery occurring only in spiking neurons and the step of distribution incorporation. To further enhance the capability of SNNs in handling temporal information, it is crucial to improve other network components besides spiking neurons. The inability of these components to handle temporal information limits the model's understanding of the video. 3D convolution is a suitable candidate that effectively preserves spatial-temporal information and is compatible with spiking neurons. Therefore, we propose replacing the 2D spatial components with 3D spatial-temporal components to facilitate the delivery of temporal information between different layers. Our framework is illustrated in Figure 4. It is important to note that SNNs can be enhanced with 3D components to better handle temporal information, and we select the SEW ResNet [7] as our baseline model.

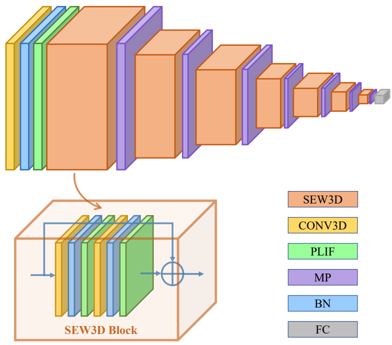


Figure 4: Architecture of *3D-SNN*, consisting of 7 spike-element-wise (*SEW*) residual blocks, each including *CONV3D*, *BN*, *PLIF*, *MaxPooling*, and *FC* layers.

# 5 Experiments

In this section, we conduct experiments on several event-based action/gesture recognition datasets to evaluate the effectiveness of our proposed SNNs. Ablation studies are carried out to quantify the effectiveness of each module.

Table 1: Comparisons of the validation accuracy with state-of-the-art methods on four datasets. *: re-implementation results based on the publicly available code.

| Methods | **FallingDetection-CeleX** | | **DVSGesture** | | **DailyAction** | **AR** |
|---|---|---|---|---|---|---|
| | Fall Detection | Action Recognition | 10-classes | 11-classes | | |
| EVENT-DRIVEN [45] | – | – | – | – | 68.3 | 55.0 |
| SPA [22] | – | – | – | – | 76.9 | – |
| Truenorth [1] | – | – | 91.8 | – | – | – |
| SLAYER [37] | – | – | 93.6 | – | – | – |
| Motion-based SNN [23] | – | – | – | 92.7 | 90.3 | 78.1 |
| PlainNet* [7] | 89.8 | 64.3 | 92.1 | 91.7 | 92.8 | 66.7 |
| Spiking ResNet* [7] | 91.8 | 70.4 | 92.1 | 90.6 | 95.8 | 67.4 |
| SEW ResNet-ADD* [7] | 93.2 | 82.7 | 97.4 | 97.2 | 98.7 | 83.0 |
| EvT [36] | – | – | 98.5 | 96.2 | – | – |
| *TS-SNN* (Ours) | 94.6 | 88.4 | **98.9** | **97.6** | **99.4** | 89.1 |
| *3D-SNN* (Ours) | **95.9** | **90.1** | 98.1 | **97.6** | 99.3 | **94.9** |

## 5.1 Datasets

**FallingDetection-CeleX** Dateset: The newly collected *FallingDetection-CeleX* dataset comprises 875 recordings of 51 subjects acting 7 different actions, namely lying down, sitting, squatting, bending, falling from a standing position (fall1), falling while getting up (fall2) and falling backward/slipping (fall3). The CeleX-V [3] event camera is used to do the recording. We select 581 clips for training and 294 clips for testing. In order to capture event sequences from various angles, each subject repeats each action from 3 different viewpoints (i.e, front view, back view, and side view). We consider two tasks in this dataset, the first one is the normal 7-class 'Action Recognition', and the other is the 'Falling Detection', which is the binary classification problem (fall down or no fall).

**DVSGesture** dataset [1]: It is a real-world gesture recognition dataset collected by the DVS128 camera with a sensor size of $128 \times 128$. It comprises 1342 recordings of 11 different actions collected on 29 individuals under 3 different lighting conditions. As suggested by the dataset paper, 23 subjects are designated as the training set, and the remaining 6 subjects are reserved as the validation set.

**DailyAction** dataset [23]: It comprises 1440 recordings of 15 subjects performing 12 different actions. A DVS camera was positioned at two different locations, each with a distinct distance and angle. The actions were recorded under two lighting conditions: *natural light* and *LED light*. Each subject performed each action under the same camera position and the same lighting condition.

**Action Recognition (AR)** dataset [28]: It has 291 recordings of 15 subjects acting 10 different actions. The recordings were captured using DAVIS cameras placed at three different distances from the subjects. Each subject performs three times for each pre-defined action.

Both the DailyAction dataset [23] and AR dataset [28] papers mention that the whole dataset is divided into the training and validation sets by 8:2, but they do not release the division details. We conduct experiments based on the 5-fold cross-validation strategy on these two datasets.

## 5.2 Implementation Details

We select the Spike-Element-Wise (SEW) ResNet [7] as our baseline model, and train the proposed *3D-SNN* and *TS-SNN* from scratch in an end-to-end manner, using standard crossentropy loss as the classification loss for optimization. The network architecture consisted of 7 blocks for all experiments. In the *3D-SNN*, the first convolutional layer implemented a $1 \times 1 \times 1$ kernel, while the following 3D convolutional layers employed a $1 \times 3 \times 3$ kernel size with 32 channels for the DVSGesture dataset and a $3 \times 3 \times 3$ kernel size with 128 channels for the other three datasets. Each max pooling layer used a $1 \times 2 \times 2$ kernel size. For *TS-SNN*, we follow the network setup in SEW ResNet [7] for all four datasets. In the experiments on the DVSGesture dataset, we choose SGD as the optimizer and set the initial learning rate to 1e-3, which was reduced with cosine annealing. Both *3D-SNN* and *TS-SNN* were trained for 192 epochs with a batch size of 16, consistent with the baseline models [7]. On the other three datasets, we use the Adam optimizer and set the initial learning rate to 1e-3, which was reduced with cosine annealing. The total number of epochs was 300, and the batch size was 8. For the *3D-SNN*, all datasets were integrated into $T = 16$ frames, with $T_{\text{train}} = 12$ frames randomly chosen for training. For the *TS-SNN*, we sample $K = 5$ frames, with a total of $N = 24$ frames and $L = 3$ segments.

## 5.3 Comparison with Existing Literature

**Results on FallingDetection-CeleX Dataset.** Table 1 shows the experimental results on *FallingDetection-CeleX* Dataset. It can be seen that our proposed methods achieve state-of-the-art performance on both the 'Fall Detection' and 'Action Recognition' setting. Notably, on the 'Action Recognition' setting, the proposed *3D-SNN* outperforms the existing method by more than 7%.

**Results on DVSGesture Dataset.** We follow the cross-subject protocol, as suggested in [1] to evaluate our frameworks and compare our *3D-SNN* and *TS-SNN* models with state-of-the-art methods, as shown in Table 1. Since there is an extra category for random movements, Table 1 shows the validation accuracy with and without including the extra category (for 11 and 10 classes classification, respectively). It can be seen that our proposed methods outperform state-of-the-art methods on both these two settings.

**Results on DailyAction Dataset.** Since there is no official training and validation set, we conduct the experiments based on the 5-fold cross-validation strategy. Our proposed methods achieve state-of-the-art performances on the DailyAction dataset, as shown in Table 1.

**Results on AR Dataset.** We compare our proposed methods with state-of-the-art methods, as shown in Table 1. We can find that our proposed *TS-SNN* and *3D-SNN* outperforms the other SNN-based methods by a large margin. Specifically, the *3D-SNN* outperforms the baseline model by around 12%. Action Recognition Dataset only contains limited training samples, the experimental results shows that our proposed *3D-SNN* can perform extremely well on the small-scale dataset.

## 5.4 Ablation Studies

In this subsection, we perform ablation studies to evaluate the impact of various convolutional kernel sizes on our *3D-SNN* and various segment and frame choices on the proposed *TS-SNN*. All experiments for ablation studies are conducted on the *FallingDetection-CeleX* Dateset under the 'Action Recognition' setting.

**1) Impact of the spatial and temporal convolutional kernel sizes:** The convolutional kernel size has a direct effect on the learned features, here we present the classification accuracy obtained using different spatial and temporal convolutional kernel sizes with *3D-SNN* architectures, as shown in Table 2. (Noted that we follow [40] to implement the even-sized kernels.)

**Temporal kernel sizes:** Row (a) in Table 2 shows results of *3D-SNN* that changing temporal kernel size. We can find that model with the temporal kernel size 3 performs best among the different temporal kernel sizes. Consequently, we choose the temporal kernel size 3 for all experiments in this paper.

**Spatial kernel sizes:** Table 2 (b) shows the results of *3D-SNN* nets that changing spatial kernel size. We can find that the best classification accuracy is 90.1%, obtained with models that have the kernel size of $3\times3\times3$ and $3\times5\times5$. Since $5\times5$ spatial convolutional kernels need much more computation cost, compared with $3\times3$ spatial kernel, we choose $3\times3\times3$ for the best trade-off between performance and efficiency for *3D-SNN* experiments in this paper.

Table 2: Comparisons of *3D-SNN* with different 3D kernel sizes.

| | | $f_t \times f_w \times f_h$ | Top-1 Acc. (%) |
|---|---|---|---|
| (a) | Temporal | $1\times3\times3$ | 88.1 |
| | | $2\times3\times3$ | 88.8 |
| | | $3\times3\times3$ | **90.1** |
| | | $4\times3\times3$ | 89.8 |
| | | $5\times3\times3$ | 88.4 |
| (b) | Spatial | $3\times2\times2$ | 87.8 |
| | | $3\times3\times3$ | **90.1** |
| | | $3\times4\times4$ | 89.1 |
| | | $3\times5\times5$ | **90.1** |
| | | $3\times6\times6$ | 89.1 |
| | | $3\times7\times7$ | 88.8 |

**2) Impact of the Number of Segments and Random Selected Frames:** As mentioned in Section 4.2, for the *TS-SNN*, we divide the event sequence into $X$ segments and randomly select $K$ frames in each segment. In this part, we study the impact of the number of segments ($X$) and the number of selected frames $K$ on the *FallingDetection-CeleX* Dateset. As shown in Table 3, sufficient performance has been achieved when $L=3, K=5$ and $L=3, K=6$. Therefore, we choose $L=3, K=5$ for the best trade-off between performance and efficiency. (Note that we do not conduct the experiments if there is only one selected frame or if all frames need to be selected.)

Table 3: Comparisons of *TS-SNN* with different numbers of segments and numbers of selected frames.

| Segments (X) | Selected Frames (K) | | | | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 |
| 3 | 86.4 | 88.1 | 87.8 | **88.4** | **88.4** | 87.4 |
| 4 | 86.4 | 87.8 | 88.1 | 88.1 | – | – |
| 6 | 87.4 | 87.8 | – | – | – | – |
| 8 | 88.1 | – | – | – | – | – |

# 6 Conclusions

In this paper, we demonstrate the potential of SNNs in combination with event-based cameras for event-based HAR. To address the limitation of SNNs in processing long-range temporal information, we propose two novel frameworks: *TS-SNN* and *3D-SNN*. The *TS-SNN* extracts long-term temporal information by dividing actions into shorter segments, while the *3D-SNN* replaces 2D spatial components with 3D ones to facilitate the delivery of temporal information between different frames. To encourage further research in event-based HAR, we create the *FallingDetection-CeleX* dataset, gathered using the high-resolution CeleX-V event camera. Our proposed frameworks surpass SOTA methods on our collected *FallingDetection-CeleX* dataset and three other neuromorphic datasets. The experimental results demonstrate the efficacy of our frameworks in processing long-range temporal information for event-based HAR. We believe that our findings will pave the way for future research in this area and provide a solid foundation for developing practical applications that address privacy concerns in HAR.

# References


[1] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017.

[2] Marco Buzzelli, Alessio Albé, and Gianluigi Ciocca. A vision-based system for monitoring elderly people at home. *Applied Sciences*, 10(1):374, 2020.

[3] Shoushun Chen and Menghan Guo. Live demonstration: Celex-v: A 1m pixel multi-mode event-based sensor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[4] Xiang Cheng, Yunzhe Hao, Jiaming Xu, and Bo Xu. Lisnn: Improving spiking neural networks with lateral interactions for robust object recognition. In *IJCAI*, pages 1519–1525, 2020.

[5] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20164–20173, 2022.

[6] Jide S Edu, Jose M Such, and Guillermo Suarez-Tangil. Smart home personal assistants: a security and privacy review. *ACM Computing Surveys (CSUR)*, 53(6):1–36, 2020.

[7] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34:21056–21069, 2021.

[8] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2661–2671, 2021.

[9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[10] Arun M George, Dighanchal Banerjee, Sounak Dey, Arijit Mukherjee, and P Balamurali. A reservoir-based convolutional spiking neural network for gesture recognition from dvs input. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

[11] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[12] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[13] Weihua He, YuJie Wu, Lei Deng, Guoqi Li, Haoyu Wang, Yang Tian, Wei Ding, Wenhui Wang, and Yuan Xie. Comparing snns and rnns on neuromorphic vision datasets: Similarities and differences. *Neural Networks*, 132:108–120, 2020.

[14] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

[15] Simone Undri Innocenti, Federico Becattini, Federico Pernici, and Alberto Del Bimbo. Temporal binary representation for event-based action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10426–10432. IEEE, 2021.

[16] Eugene M Izhikevich. *Dynamical systems in neuroscience*. MIT press, 2007.

[17] Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:424, 2020.

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

[19] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508, 2016.

[20] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 $\times$128 120 db 15$\mu$ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008.

[21] Jixin Liu, Rong Tan, Guang Han, Ning Sun, and Sam Kwong. Privacy-preserving in-home fall detection using visual shielding sensing and private information-embedding. *IEEE Transactions on Multimedia*, 23:3684–3699, 2020.

[22] Qianhui Liu, Haibo Ruan, Dong Xing, Huajin Tang, and Gang Pan. Effective aer object classification using segmented probability-maximization learning in spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1308–1315, 2020.

[23] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *IJCAI*, pages 1743–1749, 2021.

[24] Wenhe Liu, Guoliang Kang, Po-Yao Huang, Xiaojun Chang, Yijun Qian, Junwei Liang, Liangke Gui, Jing Wen, and Peng Chen. Argus: Efficient activity detection system for extended video analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 126–133, 2020.

[25] Mingqi Lu, Yaocong Hu, and Xiaobo Lu. Driver action recognition using deformable and dilated faster r-cnn with optimized region proposals. *Applied Intelligence*, 50(4): 1100–1111, 2020.

[26] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

[27] Jean-Matthieu Maro, Sio-Hoi Ieng, and Ryad Benosman. Event-based gesture recognition with dynamic background suppression using smartphone computational capabilities. *Frontiers in neuroscience*, 14:275, 2020.

[28] Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois Knoll. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in neurorobotics*, 13:38, 2019.

[29] Daniel Neil and Shih-Chii Liu. Effective sensor fusion with event-based sensors and deep network architectures. In *2016 IEEE international symposium on circuits and systems (ISCAS)*, pages 2282–2285. IEEE, 2016.

[30] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. *Advances in neural information processing systems*, 29, 2016.

[31] Garrick Orchard, Cedric Meyer, Ralph Etienne-Cummings, Christoph Posch, Nitish Thakor, and Ryad Benosman. Hfirst: A temporal approach to object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):2028–2040, 2015.

[32] Nicolas Perez-Nieves and Dan Goodman. Sparse spiking gradient descent. *Advances in Neural Information Processing Systems*, 34:11795–11808, 2021.

[33] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2010.

[34] Mamshad Nayeem Rizve, Ugur Demir, Praveen Tirupattur, Aayush Jung Rana, Kevin Duarte, Ishan R Dave, Yogesh S Rawat, and Mubarak Shah. Gabriella: An online system for real-time activity detection in untrimmed security videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4237–4244. IEEE, 2021.

[35] Isidoros Rodomagoulakis, Nikolaos Kardaris, Vassilis Pitsikalis, E Mavroudi, Athanasios Katsamanis, Antigoni Tsiami, and Petros Maragos. Multimodal human action recognition in assistive human-robot interaction. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2702–2706. IEEE, 2016.

[36] Alberto Sabater, Luis Montesano, and Ana C Murillo. Event transformer. a sparse-aware solution for efficient event data processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2677–2686, 2022.

[37] Sumit B Shrestha and Garrick Orchard. Slayer: Spike layer error reassignment in time. *Advances in neural information processing systems*, 31, 2018.

[38] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[39] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018.

[40] Shuang Wu, Guanrui Wang, Pei Tang, Feng Chen, and Luping Shi. Convolution with even-sized kernels and symmetric padding. *Advances in Neural Information Processing Systems*, 32, 2019.

[41] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1311–1318, 2019.

[42] Yujie Wu, Rong Zhao, Jun Zhu, Feng Chen, Mingkun Xu, Guoqi Li, Sen Song, Lei Deng, Guanrui Wang, Hao Zheng, et al. Brain-inspired global-local learning incorporated with neuromorphic computing. *Nature Communications*, 13(1):1–14, 2022.

[43] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[44] Rong Xiao, Huajin Tang, Yuhao Ma, Rui Yan, and Garrick Orchard. An event-driven categorization model for aer image sensors using multispike encoding and learning. *IEEE transactions on neural networks and learning systems*, 31(9):3649–3657, 2019.

[45] Rong Xiao, Huajin Tang, Yuhao Ma, Rui Yan, and Garrick Orchard. An event-driven categorization model for aer image sensors using multispike encoding and learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3649–3657, 2020. doi: 10.1109/TNNLS.2019.2945630.

[46] Yannan Xing, Gaetano Di Caterina, and John Soraghan. A new spiking convolutional recurrent neural network (scrnn) with applications to event-based hand gesture recognition. *Frontiers in neuroscience*, 14:1143, 2020.

[47] Friedemann Zenke and Emre O Neftci. Brain-inspired learning on neuromorphic substrates. *Proceedings of the IEEE*, 109(5):935–950, 2021.