# UCF-Crime-DVS: A Novel Event-Based Dataset for Video Anomaly Detection with Spiking Neural Networks

**Yuanbin Qian[1*], Shuhan Ye[1*], Chong Wang[1,2†], Xiaojie Cai[1], Jiangbo Qian[1,2], Jiafei Wu[3]**

[1]Faculty of Electrical Engineering and Computer Science, Ningbo University, China
[2]Merchants' Guild Economics and Cultural Intelligent Computing Laboratory, Ningbo University, China
[3]Department of Electrical and Electronic Engineering, The University of Hong Kong
{2311100301, 216002718, wangchong, 2211100083, qianjiangbo}@nbu.edu.cn, jcjiafeiwu@gmail.com

## Abstract

Video anomaly detection plays a significant role in intelligent surveillance systems. To enhance model's anomaly recognition ability, previous works have typically involved RGB, optical flow, and text features. Recently, dynamic vision sensors (DVS) have emerged as a promising technology, which capture visual information as discrete events with a very high dynamic range and temporal resolution. It reduces data redundancy and enhances the capture capacity of moving objects compared to conventional camera. To introduce this rich dynamic information into the surveillance field, we created the first DVS video anomaly detection benchmark, namely UCF-Crime-DVS. To fully utilize this new data modality, a multi-scale spiking fusion network (MSF) is designed based on spiking neural networks (SNNs). This work explores the potential application of dynamic information from event data in video anomaly detection. Our experiments demonstrate the effectiveness of our framework on UCF-Crime-DVS and its superior performance compared to other models, establishing a new baseline for SNN-based weakly supervised video anomaly detection.

**Dataset and Code** —
https://github.com/YBQian-Roy/UCF-Crime-DVS

## Introduction

Video anomaly detection (VAD) is a crucial research direction in the fields of computer vision and machine learning, which plays a significant role in intelligent video surveillance system (Zhou, Yu, and Yang 2023). For VAD tasks, content-rich datasets are effective in evaluating the strengths and weaknesses of algorithms and models. Benchmark datasets help define the scope of problems that can be solved. Some of the common publicly available benchmark datasets for VAD include UCSD-Peds (Li, Mahadevan, and Vasconcelos 2013), Avenue (Lu, Shi, and Jia 2013), Street Scene (Ramachandra and Jones 2020), Shanghai Tech (Luo, Liu, and Gao 2017), TAD (Lv et al. 2021), and UCF-Crime (Sultani, Chen, and Shah 2018), which cover various monitoring scenarios and anomalous events. Generally,

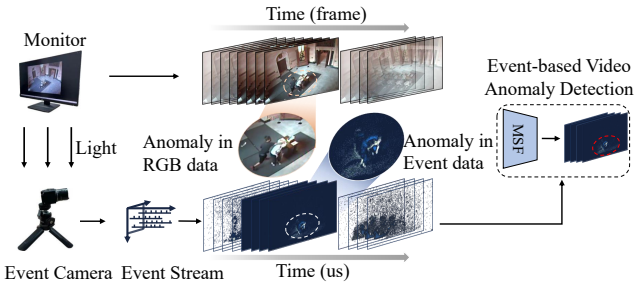


Figure 1: The overview of our contributions.

these datasets are first processed through a feature extractor to obtain RGB features or optical flow features. RGB features capture the appearance information of the video, while optical flow features focus on the motion information.

Recently, dynamic vision sensors (DVS) (Lichtsteiner, Posch, and Delbruck 2008; Brandli et al. 2014), also known as event cameras, have garnered significant attention due to their high dynamic range, high temporal resolution, and low latency. DVS is a bionic visual sensor inspired by human retinal peripheral neurons. It uses a difference-based sampling model to generate event data only when pixel brightness changes above a threshold. Unlike traditional images, event streams encode visual information as discrete events, dramatically reducing data redundancy and preserving temporal characteristics. This efficient information processing has enabled event cameras to capture moving objects in the frame better than conventional cameras and reduce system-level power consumption by up to 100 times (Delbrück et al. 2010; Posch, Matolin, and Wohlgenannt 2010). However, despite their advantages, event cameras have not yet been applied to the field of VAD. Therefore, we introduced the first DVS dataset in this field using an event camera, called UCF-Crime-DVS, to explore the potential of it in VAD.

However, Artificial Neural Networks (ANNs) do not handle event streams well due to the discrete nature of event data. Unlike ANNs, Spiking Neural Networks (SNNs) receive event format data as input and use discrete, binary spike signals, leading to a natural advantage in handling event streams (Chen et al. 2023). Therefore, to better utilize event data in the field of VAD, this paper introduces a fully SNN-based VAD framework called MSF. Given the unique

*These authors contributed equally.
†Corresponding author: Chong Wang.

| Dataset | Sensors | Class | Resolution | Sec Per Example | Object |
|---|---|---|---|---|---|
| N-Caltech101(Orchard et al. 2015) | ATIS | 101 | $240 \times 180$ | 0.3s | images |
| N-MNIST(Orchard et al. 2015) | ATIS | 10 | $28 \times 28$ | 0.3s | images |
| CIFAR10-DVS(Li et al. 2017) | DAVIS128 | 10 | $128 \times 128$ | - | images |
| N-ImageNet(Kim et al. 2021) | Samsung-Gen3 | 1000 | $346 \times 260$ | - | images |
| ES-ImageNet(Lin et al. 2021) | - | 1000 | $224 \times 224$ | - | images |
| DVS-Gesture(Amir et al. 2017) | DAVIS128 | 11 | $128 \times 128$ | 6s | action |
| N-CARS(Sironi et al. 2018) | ATIS | 2 | $128 \times 128$ | 0.1s | cars |
| ASL-DVS(Bi et al. 2019) | DAVIS240 | 24 | $346 \times 260$ | 0.1s | hand |
| ASLAN-DVS(Bi et al. 2020) | DAVIS240c | 432 | $240 \times 180$ | - | action |
| HMDB-DVS(Bi et al. 2020) | DAVIS240c | 51 | $240 \times 180$ | 19s | action |
| UCF101-DVS(Bi et al. 2020) | DAVIS240c | 101 | $240 \times 180$ | 25s | action |
| PAF(Miao et al. 2019) | DAVIS346 | 10 | $346 \times 260$ | 5s | action |
| DailyAction(Liu et al. 2021) | DAVIS346 | 12 | $346 \times 260$ | 5s | action |
| HARDVS(Wang et al. 2024) | DAVIS346 | 300 | $346 \times 260$ | 5-10s | action |
| Bullying10K(Dong et al. 2024) | DAVIS346 | 10 | $346 \times 260$ | 2-20s | action |
| **UCF-Crime-DVS (Ours)** | IMX636 | 14 | $\mathbf{1280 \times 720}$ | **avg 242s** | **anomaly** |

Table 1: Overview of various DVS datasets.

dynamics and temporal complexity of event data, effectively processing this complexity is important. TIM (Shen et al. 2024) enhances the spiking self-attention (SSA) mechanism's ability to handle these challenges. Consequently, our MSF incorporates TIM to improve the model's temporal processing of event data.

To the best of our knowledge, this work pioneers the exploration of applying event data to VAD. The overview of our work is illustrated in Figure 1. First, we constructed an event-based dataset for VAD. With this dataset, we then present the MSF framework, a fully spiking neural network architecture designed to better detect anomalous events from event streams. Overall, our contributions can be summarized as follows:

- We present the first large DVS dataset for VAD, in order to apply the rich dynamic information and high temporal resolution of DVS in VAD.
- We propose a multi-scale SNN-based framework for DVS-based VAD. The Temporal Interaction Module (TIM) is innovatively incorporated in the convolution-based SNNs framework to enhance the integration of spiking features, demonstrating its effectiveness on other time-series tasks.

## Related Works

### Event Camera Applications

Event cameras have been widely used in computer vision applications. For example, TEF (Han et al. 2023) reconstructs image signals by converting the high temporal resolution of the event signals into precise radiance values. SAN (Zhang et al. 2023) allows flexible input spatial scaling and uses self-supervised fine-tuning to enhance generalization performance for removing motion blur from images. STNet (Zhang et al. 2022) dynamically extracts and fuses information from temporal and spatial domains for single-target tracking. ExACT (Zhou et al. 2024) introduces a novel approach to event-based action recognition by employing a cross-modal conceptualization. Although event cameras have been applied in many areas of computer vision, they have not yet been utilized in VAD. Therefore, our work explores this possibility.

### Weakly Supervised Video Anomaly Detection

Our work is a weakly supervised video anomaly detection (WSVAD) task. The mainstream approach to it is multi-instance learning (MIL), proposed by (Sultani, Chen, and Shah 2018). Specifically, MIL treats each video as a "bag" and divides each video into equal-length, non-overlapping segments called instances. All instances in normal videos are called positive bags, while those that contain at least one abnormal instance are called negative bags, representing abnormal videos. In MIL, learning is performed by decreasing the predicted anomaly score for each instance in the positive bag and increasing the score only for the instance with the largest anomaly score in the negative bag. Overall, WSVAD can be summarized in three stages: 1) each video is segmented into multiple clips, and features are extracted by a pre-trained encoder; 2) anomaly scores are generated using a multilayer perceptron (MLP); 3) the model is optimized using the MIL framework.

### Spiking Neurons

Since event cameras record the visual input as asynchronous discrete events, they are inherently suitable to cooperate with SNNs. Spiking neurons in forward propagation can be summarized in three steps: charge, fire, and reset (Fang et al. 2021a). In this paper, we choose the leaky integrate-and-fire (LIF) neuron model (Gerstner et al. 2014), which is widely adopted in SNNs due to its simplicity and ability to capture key aspects of neuronal dynamics. The dynamic model of LIF can be written in the following form:

$$\mathbf{u}^{t+1,l} = \tau \mathbf{u}^{t,l} + \mathbf{W}^l \mathbf{o}^{t,l-1}, \tag{1}$$

$$\mathbf{o}^{t,l} = \Theta(\mathbf{u}^{t,l} - V_{th}), \tag{2}$$

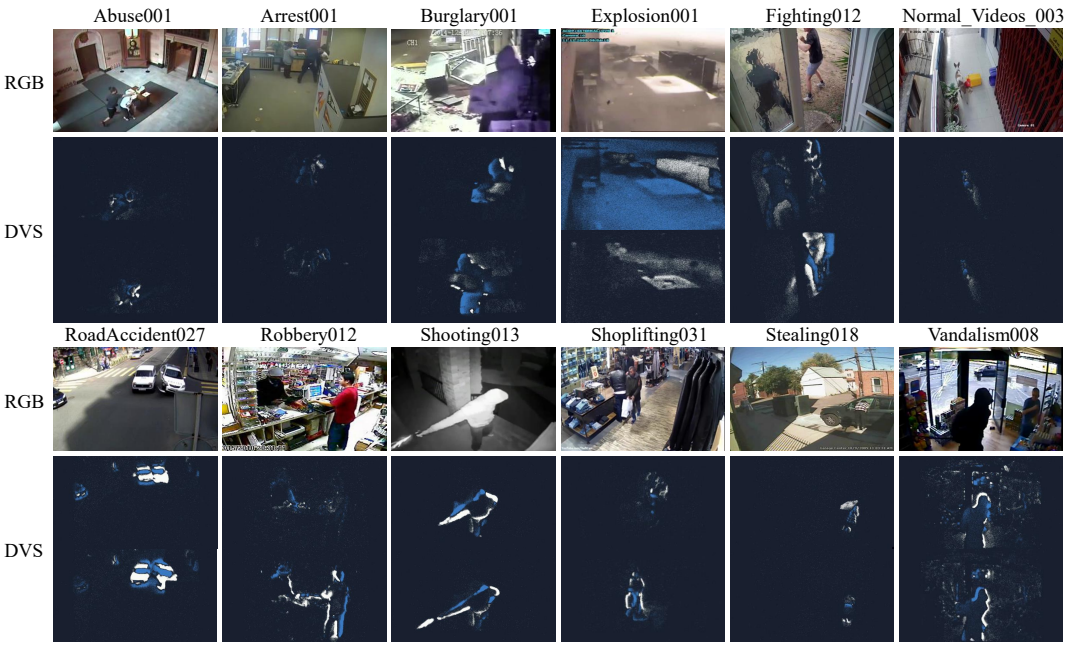


Figure 2: Presentation of our dataset and comparison between the DVS and RGB version of UCF-Crime.

$$\mathbf{u}^{t+1,l} = \tau \mathbf{u}^{t,l} \cdot (1 - \mathbf{o}^{t,l}) + \mathbf{W}^{l} \mathbf{o}^{t+1,l-1}, \quad (3)$$

where $\tau$ is leaky factor and $u^{t,l}$ denotes membrane potential of the neurons in layer $l$ at time step $t$. $\mathbf{W}^{l}$ and $\mathbf{o}^{l}$ represent the weight parameters and the fired spikes, respectively. $\Theta$ denotes Heaviside step function. When $\mathbf{u}^{t,l} \geq V_{th}$ equals to one, otherwise equals to zero. The membrane potential accumulates with the input until a given threshold $V_{th}$ is exceeded, then the neuron delivers a spike and the membrane potential $\mathbf{u}^{t,l}$ is reset to zero.

## UCF-Crime-DVS Dataset

For VAD, datasets are as fundamental as models. In our paper, we construct the first event-based VAD dataset, named UCF-Crime-DVS. Our dataset contains 1900 event streams across 13 anomaly classes, aligned with the original UCF-Crime dataset (Sultani, Chen, and Shah 2018). It includes 1610 training sets with video-level labels and 290 test sets with frame-level labels, maintaining an equal number of normal and abnormal videos in both. Table 1 compares the parameters with other DVS dataset, highlighting that our dataset has a high resolution of 1280×720 and an average duration of 242 seconds per video, totaling 128 hours. This far exceeds the specifications of other DVS datasets. Next, we will demonstrate the characteristics of the dataset and provide a detailed description of the dataset construction process.

### Characteristics of Event Data

Unlike pixel points in RGB video, which has three channels $(red, green, blue)$, event data consists of only two channels $(OFF, ON)$. Specifically, each event can be represented by $e = (x, y, p, t)$, where $(x, y)$ represents the position, $p \in \{0, 1\}$ indicates the polarity, and $t$ represents the timestamp (microsecond, $\mu$s). Events where the brightness increases above the threshold are called $ON$ events, while those where the brightness decreases are called $OFF$ events. As mentioned above, the event camera uses a difference-based sampling model and a threshold mechanism to generate events. This mechanism allows event cameras to capture faster moving object and more dynamic information than RGB cameras while ignoring most static information. As shown in Figure 2, the dynamic subjects in our dataset are clearly presented, whereas the static background is barely visible. Additionally, small events at the frame edges, such as in Shoplifting031 and Stealing018, can be captured by the event camera.


Figure 3: The final shooting environment setup.

### Dataset Construction

**Pre-Production Stage.** First of all, we prepared an event camera with a resolution of 1280 × 720 and IMX636 sensors provided by Prophesee, and a 32" 4K monitor to play the original UCF-Crime dataset. The dataset was captured in a

light-free environment, where the only light perceived by the event camera came from the monitor playing the videos.

**Dataset Pre-Processing Stage.** We combined the videos in the original dataset by class into single long videos for playback and recorded the number of frames in each video.

**Dataset Shooting Stage.** Metavision SDK is used to control the event camera. We adjusted the aperture and focus distance to capture sharp images. To reduce background noise, we fine-tuned the bias settings while following the event rate and the display to assess the noise impact. The final shooting setup is presented in Figure 3.

**Dataset Post-Processing Stage.** When the event dataset is recorded, we slice the long event stream by the length of each video segment, ensuring alignment with the original dataset. Since the discrete event data cannot be easily processed by the downstream networks, it need to be converted into a more usable format. The mainstream method integrates the event data into event frames based on the number of event frames or duration for downstream tasks. Similarly, we merged each event stream into event frames at designated time intervals.

All events $e$ in every 533,328 $\mu$s (corresponding to 16 video frames) are integrated into an event frame $E_j$ which represents $j$-th event frame. Define $e_{\Delta t} = (x, y, p)$ as the event in $\Delta t$, where $\Delta t = t_{j_r} - t_{j_l}$. The process of integrating event can be expressed as:

$$E_j(x, y, p) = \sum_{t=t_{j_l}}^{t_{j_r}-1} \mathbf{1}(e_{\Delta t} = (x_t, y_t, p_t)), \tag{4}$$

here, $E_j(x, y, p)$ denotes the pixel value at position $(x, y, p)$ with which is integrated from the event data within the specified time interval $[t_{j_l}, t_{j_r})$ and $\mathbf{1}$ is an indicator function that equals 1 only when $e_{\Delta t} = (x_t, y_t, p_t)$.

# Methods

To effectively process the binary event streams dataset, we propose a multi-scale spiking fusion (MSF) network for WSVAD. Benefiting from the high temporal resolution and rich dynamic details of event data, the proposed multi-scale spiking fusion module can efficiently exploit temporal features. The complete structure of MSF is shown in Figure 4.

## Problem Statement

Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ represent the training set containing $n$ event stream videos from the proposed UCF-Crime-DVS dataset, and $\mathbf{T} = \{t_i\}_{i=1}^n$ denote the temporal duration, where $t_i$ is the event frame number of the $i$-th event stream. Additionally, we use $\mathbf{Y} = \{y_i\}_{i=1}^n$, where $y_i = \{0, 1\}$, to represent the video anomaly label set. In the testing stage, the anomaly score vector for $i$-th video is defined as $\mathbf{s}_i = \{s^j\}_{j=1}^t$, where $s^j = \{0, 1\}$, and $s^j$ is anomaly score of the $j$-th event clip.

## Feature Extraction

Most VAD tasks start with feature extraction. We use Hardvs dataset (Wang et al. 2024), a large event-based action recognition dataset, to pre-train a Spikingformer (Zhou et al. 2023), which serves as our feature extractor. The features of UCF-Crime-DVS are then extracted using the pre-trained Spikingformer. After that, we obtain the event stream feature $\mathbf{F}$ with dimensions $t \times D$ from the training video $\mathbf{x}$, where $D$ is the dimension of clip features. According to the multi-instance learning principle, the feature $\mathbf{F}$ is fed into our MSF.

## Multi-Scale Spiking Fusion

When dealing with event data, particularly for VAD, it is crucial to efficiently extract and retain temporal features while discovering temporal dependencies. Our proposed multi-scale spiking fusion module (MSF) captures both multi-resolution local spiking dependencies (light green block in Figure 4) within individual clip, and global spiking dependencies (light yellow block in Figure 4) between event clips. Finally, these dependencies are seamlessly integrated based on the unique characteristics of the spiking feature (light blue block in Figure 4).

**Local Spiking Feature.** MSF uses pyramidal dilated convolution $\{\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3\}$ over the temporal domain to learn multi-scale representations of event clips. It learns the multi-scale spiking features from the feature $\mathbf{F} = \{\mathbf{f}_d\}_{d=1}^D$. Given the feature $\mathbf{f}_d \in \mathbb{R}^t$, the one-dimensional dilated convolution operation is performed using the kernel $\mathbf{W}_{p,d} \in \mathbb{R}^\omega$ with $p \in \{1, ..., D/4\}$, $d \in \{1, ..., D\}$, and $\omega$ indicating the filter size, which is defined as:

$$\mathbf{f}_p = \sum_{d=1}^{D} \mathbf{W}_{p,d} * \mathbf{f}_d, \tag{5}$$

where $*$ denotes the dilated convolution operator, and $\mathbf{f}_p \in \mathbb{R}^t$ represents the output feature after applying dilated convolution in the time dimension. The features $\mathbf{F}_p \in \mathbb{R}^{t \times D/4}$that have been concatenated by $\mathbf{f}_p$ are then passed through spike neurons to obtain the spiking features:

$$\mathbf{F}^{\mathrm{P}} = Lif(\mathbf{F}_p), \tag{6}$$

where $Lif$ is the leaky integrate-and-fire spike neuron.

**Global Spiking Feature.** Despite of the local temporal dependencies, global ones are also important. We introduce a lightweight SpikingGCN to further capture the temporal dependencies across different event clips, which is shown in the yellowish green block in Figure 4. Our global temporal extraction module first downscales the features from $\mathbf{F} \in \mathbb{R}^{t \times D}$ to $\mathbf{F}^c \in \mathbb{R}^{t \times D/4}$ with $\mathbf{F}^c = Conv_{1\times1}(\mathbf{F})$. SpikingGCN then models global temporal dependencies of spiking feature in terms of feature similarity and relative distance.

Feature similarity branch generates the adjacency matrix $\mathbf{M}_{sim}$ for SpikingGCN using event frame-wise cosine similarity method, which is denoted as follows,

$$\mathbf{M}^{sim} = \frac{\mathbf{F}^c \mathbf{F}^{c\top}}{\|\mathbf{F}^c\|_2 \cdot \|\mathbf{F}^c\|_2}. \tag{7}$$

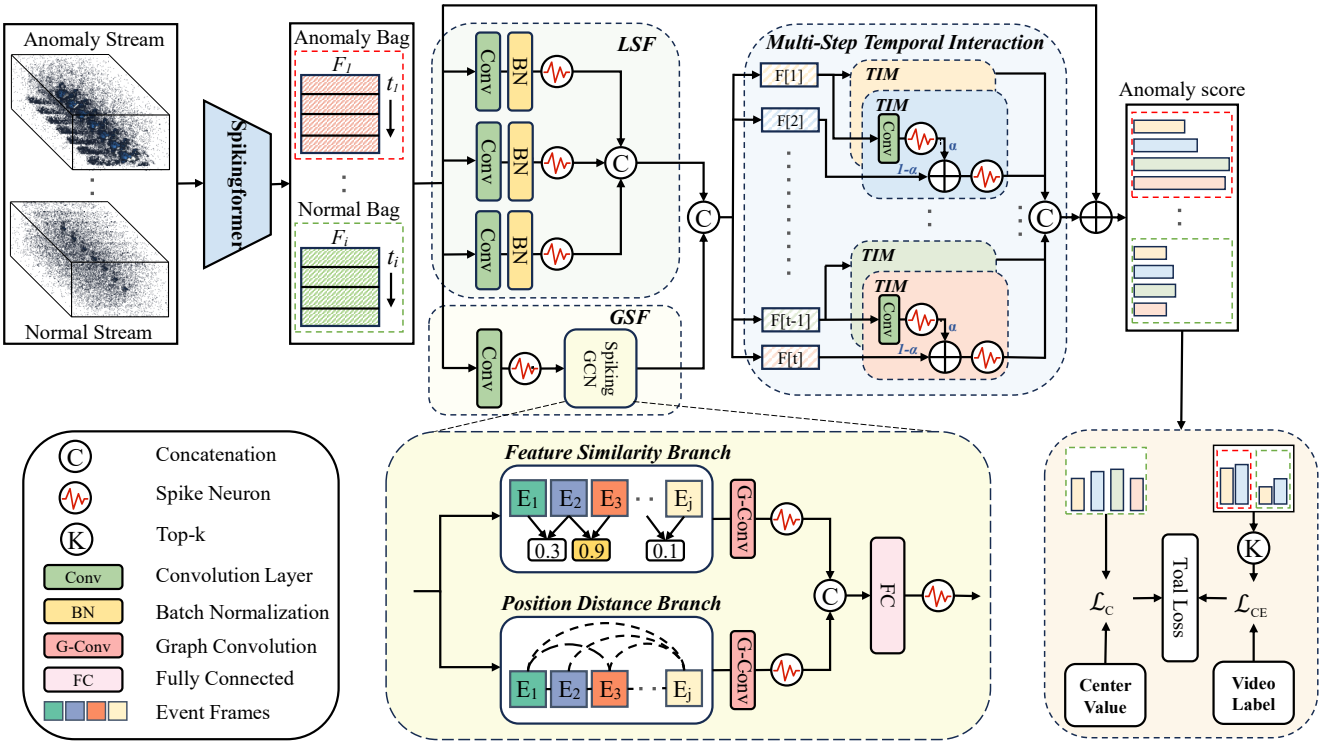


Figure 4: The framework of our proposed MSF. The LSF and GSF represents the local and global spiking feature extractor module, respectively. $\mathcal{L}_{\mathrm{CE}}$ denotes cross-entropy loss, and $\mathcal{L}_{\mathrm{C}}$ denotes center loss.

We employ the position distance branch to capture long-distance relationships between objects or scenes by measuring their positional differences across event frames, as illustrated below:

$$\mathbf{M}^{dis}(i,j) = \frac{-|i-j|}{\sigma}, \tag{8}$$

it means the proximity between event frames $i$ and $j$ depends solely on their relative positions in time, independent of other factors. The hyperparameter $\sigma$ is used to adjust the degree of influence.

Overall, the modified SpikingGCN can be summarized as follows:

$$\mathbf{F}^{\mathrm{G}} = Lif\left(\left[Soft(\mathbf{M}^{sim}); Soft(\mathbf{M}^{dis})\right]\mathbf{F}^{c}\mathbf{W}\right), \tag{9}$$

where $\mathbf{W}$ is the unique learnable weight used to transform the input feature space into another feature space. $Soft$ indicates the Softmax normalization, which is used to ensure the sum of each row of $\mathbf{M}^{sim}$ and $\mathbf{M}^{dis}$ equals to one.

**Multi-Scale Spiking Interaction.** We use residual concatenation to prevent features from being over-smoothed and concatenate global spiking features with local spiking features, which can be describe as,

$$\bar{\mathbf{F}} = [\mathbf{F}^{(l)}]_{l\in L} \in \mathbb{R}^{t\times D}, \tag{10}$$

where $L = \{\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3, \mathrm{G}\}$. $\mathbf{F}^{\mathrm{P}}$ and $\mathbf{F}^{\mathrm{G}}$ refer to the learned local and global temporal features respectively.

As previously mentioned, event data possesses unique dynamics and temporal intricacies, whereas the membrane potentials of spiking neurons exhibit a cumulative nature. Therefore, the extracted temporal information initially manifests as membrane potentials rather than spikes. Consequently, traditional ANN-based temporal learning methods, such as MTN (Tian et al. 2021), fail to effectively integrate the multi-scale features of event clips, leading to substantial under-utilization of information from different time steps. To exploit the hidden information across various time steps, we employ the Temporal Interaction Module (TIM) (Shen et al. 2024) to fuse historical spike information with current spike information. The hyperparameter $\alpha$ is used as a weight parameter, allowing the model to balance the combination of historical states and current inputs during computation. This can be mathematically expressed by the following equation:

$$\mathbf{F}^{\mathrm{TIM}} = \alpha Conv(\mathbf{F}^{\mathrm{TIM}}[t-1]) + (1-\alpha)\bar{\mathbf{F}}[t]. \tag{11}$$

TIM demonstrates a dual mechanism for temporal information processing: immediate feature extraction and historical state integration. This approach not only extracts key features from the current input but also effectively utilizes the implicit state information from previous time steps. This design achieves an organic combination of short-term and long-term dependencies, enabling the model to capture the complex dynamics in event data.

| Methods | Architecture | Supervision | AUC(%) | FAR(%) |
|---|---|---|---|---|
| Sultani et al. (Sultani, Chen, and Shah 2018) | ANNs | Weakly-supervised | 55.56 | 8.69 |
| 3C-Net (Narayan et al. 2019) | ANNs | Weakly-supervised | 59.22 | 9.50 |
| AR-Net (Wan et al. 2020) | ANNs | Weakly-supervised | 60.71 | 8.51 |
| Wu et al. (Wu et al. 2020) | ANNs | Weakly-supervised | 58.58 | 34.35 |
| RTFM (Tian et al. 2021) | ANNs | Weakly-supervised | 52.67 | 13.19 |
| TSA (Joo et al. 2023) | ANNs | Weakly-supervised | 51.86 | 22.36 |
| SEW-ResNet (Fang et al. 2021a) | SNNs | Weakly-supervised | 53.99 | 11.79 |
| PLIF (Fang et al. 2021b) | SNNs | Weakly-supervised | 54.74 | 9.17 |
| baseline(Zhou et al. 2023) | SNNs | Weakly-supervised | 62.78 | 11.52 |
| **MSF(Ours)** | SNNs | Weakly-supervised | **65.01** | **3.27** |

Table 2: AUC and FAR of the proposed method against other methods on UCF-Crime-DVS. These methods are adapted to our achitecture and re-trained on the UCF-Crime-DVS.

| LSF | GSF | TIM | AUC(%) | FAR(%) |
|---|---|---|---|---|
| - | - | - | 62.78 | 11.52 |
| ✓ | - | - | 62.44 | 5.06 |
| - | ✓ | - | 60.32 | 6.80 |
| - | - | ✓ | 50.06 | **0.68** |
| ✓ | ✓ | - | 55.69 | 7.27 |
| ✓ | - | ✓ | 64.39 | 2.80 |
| - | ✓ | ✓ | 64.07 | 5.13 |
| ✓ | ✓ | ✓ | **65.01** | 3.27 |

Table 3: Ablation study for different module.

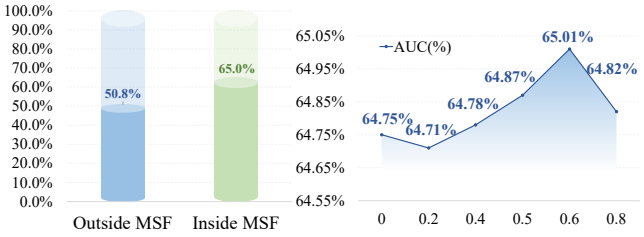


Figure 5: Ablation study for TIM. Left: Performance comparison with different position of TIM. Right: Performance comparison with different $\alpha$ values.

| $\tau$ | 0.2 | 0.25 | 0.4 | 0.5 | 0.625 | 0.8 |
|---|---|---|---|---|---|---|
| AUC(%) | 63.99 | 63.96 | 64.37 | 64.78 | **65.01** | 64.00 |

Table 4: Performance comparison with different time constant $\tau$ on UCF-Crime-DVS.

**Anomaly Scorer.** After MSF, a fully connected (FC) layer and a sigmoid function are employed as an anomaly scorer to generate the anomaly score vector $\mathbf{s_i}$:

$$\mathbf{s_i} = Sigmoid(FC(\mathbf{F}^{\mathrm{TIM}})). \tag{12}$$

## Loss Function

The classic dynamic multiple-instance learning (DMIL) loss and center loss are implemented for our proposed MSF.

**DMIL Loss.** The DMIL loss is to enlarge the inter-class distance of instances, which can be represented as follows:

$$\mathcal{L}_{\mathrm{DMIL}} = \frac{1}{k_i} \sum_{s_i^j \in \mathbf{S}_i} [-y_i log(s_i^j) + (1-y_i) log(1-s_i^j)], \tag{13}$$

where $\mathbf{s}_i^j$ is descending sorted anomaly score vector of the $i$-th video, $\mathbf{S}_i = \{s_i^j \mid j = 1, 2, ..., k_i\}$ consists of top-$k_i$ elements in $s_i$ and $y_i = \{0, 1\}$ is the video anomaly label.

**Center Loss.** The center loss used for anomaly score regression collects the anomaly scores of normal event clips, reducing the intra-class distance. It can be represented as,

$$\mathcal{L}_c = \begin{cases} \frac{1}{t_i} \sum_{j=1}^{t_i} \left\| s_i^j - c_i \right\|_2^2, & \text{if } y_i = 0 \\ 0, & \text{otherwise} \end{cases}, \tag{14}$$

$$c_i = \frac{1}{t_i} \sum_{j=1}^{t_i} s_i^j, \tag{15}$$

where $c_i$ is the center of anomaly score vector $s_i$. Overall, the total loss function of our MSF model is given by:

$$\mathcal{L} = \mathcal{L}_{\mathrm{DMIL}} + \lambda \mathcal{L}_c. \tag{16}$$

# Experiments

We validated our UCF-Crime-DVS dataset and MSF framework using a VAD task. Additionally, we tested the ability of each module with ablation experiments.

## Experiments Setup

**Training Dataset.** We use our UCF-Crime-DVS dataset to test and verify our proposed method. Our UCF-Crime-DVS dataset is aligned with the UCF-Crime dataset, covering 13 classes of anomalies in 1,610 training videos with video-level labels and 290 test videos with frame-level labels.

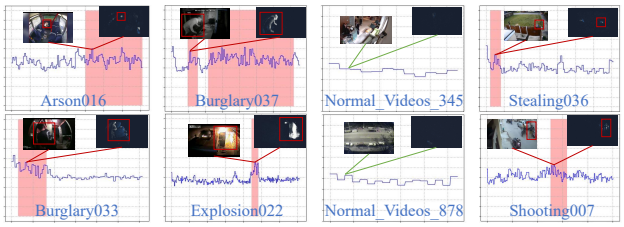


Figure 6: Anomaly scores of our methods on UCF-Crime-DVS. Pink areas indicate the manually labelled abnormal events, purple lines represent the anomaly score and red boxes point out abnormal events on the screen.

**Training Details.** Following (Sultani, Chen, and Shah 2018), each event stream is divided into non-overlapped clips. Empirically, we set $k = 4$ for our dataset. We use the Adam optimizer with a weight decay of 0.0005, and a learning rate of 0.0001. For $\sigma$ in Eq.8 and $\lambda$ in Eq.16, we set them as 1 and 20, respectively. Each batch contains 60 samples, split equally between normal and abnormal video sequences, which are randomly selected from the training set. The models for conducting experiments are implemented based on Pytorch (Paszke et al. 2019), SpikingJelly (Fang et al. 2023) and a server with single RTX4090 GPU.

**Evaluation Metrics.** We use two standardized performance metrics to evaluate the anomaly detection capability of the model: Area Under of Curve (AUC) of the frame-level Receiver Operating Characteristics (ROC) and False Alarm Rate (FAR) with a threshold 0.5. The combined assessment of these two metrics not only reflects the overall discriminative ability of the model, but also its reliability and stability in real-world application scenarios.

## Performance Analysis

Table 2 presents a comparison of our method against other methods on UCF-Crime-DVS dataset. It can be seen that compared classical VAD frameworks do not perform well on this dataset. Some methods have a FAR of more than 20%, suggesting they are unable to process event data effectively. Other SNN-based architectures with deeper network layers also fail to achieve a high AUC and low FAR at the same time, indicating that simply increasing the network complexity does not improve the VAD performance. Our MSF, on the other hand, achieves an AUC of 65.01% for anomaly detection, along with a FAR of only 3.27%, which has 3% more AUC and 8% lower FAR than our baseline. It establishes a new baseline for event-based WSVAD.

## Ablation Study

A series of ablation studies presented in Table 3 demonstrate that optimal performance is achieved when all three modules are combined. In contrast, the performance of the LSF-GSF combination, as well as each module individually, is suboptimal. This can be attributed to the fact that both LSF and GSF expand feature representations in the temporal domain, with the LSF lacking interconnections and the GSF smoothing features, which imposes a challenge for anomaly localization. However, integrating the TIM module with LSF and GSF significantly improves performance, highlighting the TIM module's critical role in effectively integrating information across time steps.

**Ablations for TIM.** As shown in the left of Figure 5, we conducted ablation experiments to examine the impact of TIM placement. The results reveal a more than 10% accuracy difference between the two placements, indicating that the optimal placement of TIM is within the MSF module. Integrating TIM within MSF enables seamless fusion of temporal features across multiple time steps, ensuring accurate capture of temporal dependencies and improving anomaly detection performance. Additionally, as seen in the right of Figure 5, any non-zero value for $\alpha$ yields better results than setting $\alpha$ to zero. MSF achieves its best performance when $\alpha$ is set to 0.6, indicating that the introduction of temporal interaction significantly enhances performance.

**Ablations for Time Constant.** A smaller time constant $\tau$ results in more leakage of the membrane potential over time, potentially leading to a loss of temporal information. To optimize the model, we performed ablation experiments on the time constants $\tau$. Table 4 shows that the model's ability to detect anomalies initially increases with $\tau$, reaching a peak accuracy of 65.01% when $\tau$ is set to 0.625. However, when $\tau$ exceeds 0.625, AUC begins to decline, dropping further at 0.8. This suggests that an excessively large $\tau$ reduces the model's temporal memory capacity. Therefore, it is indispensable to select an appropriate $\tau$ through experimentation.

## Visualization

A set of visualization are presented in Figure 6. Noting that certain scene transitions and the opening or closing credits exhibit similar characteristics to the explosion events, e.g. flickering visuals and a surge in event occurrences. It makes detecting explosion events in our dataset very challenging. However, our model still successfully recognizes explosion events such as Explosion022, highlighting its robustness. Additionally, for subtle anomalous events like stealing, which are difficult to detect visually, our model is able to identify these weak anomalies to some extent, as illustrated by the case of Stealing036. Although the visualized anomaly scores do not consistently exceed the anomaly threshold in anomalous segments, this is because some anomalous events include relatively stationary segments that did not trigger DVS, resulting in partial loss of these events.

# Conclusion

In this paper, we present the first event-based VAD dataset and introduce the MSF framework for SNN-based VAD. Extensive experiments demonstrate that our method outperforms others on UCF-Crime-DVS, highlighting its potential for real-world applications. While our method has not yet achieved such high accuracy of traditional approaches on RGB dataset, it offers a fresh perspective on VAD and lays the foundation for future research.

## Acknowledgments

This work was supported by the Ningbo Municipal Natural Science Foundation of China (No. 2022J114), National Natural Science Foundation of China (No. 62271274), Ningbo S&T Project (No.2024Z004) and Ningbo Major Research and Development Plan Project (No.2023Z225).

## References

Amir, A.; Taba, B.; Berg, D.; Melano, T.; McKinstry, J.; Di Nolfo, C.; Nayak, T.; Andreopoulos, A.; Garreau, G.; Mendoza, M.; et al. 2017. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7243–7252.

Bi, Y.; Chadha, A.; Abbas, A.; Bourtsoulatze, E.; and Andreopoulos, Y. 2019. Graph-based object classification for neuromorphic vision sensing. In *Proceedings of the IEEE/CVF international conference on computer vision*, 491–501.

Bi, Y.; Chadha, A.; Abbas, A.; Bourtsoulatze, E.; and Andreopoulos, Y. 2020. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29: 9084–9098.

Brandli, C.; Berner, R.; Yang, M.; Liu, S.-C.; and Delbruck, T. 2014. A 240× 180 130 db 3 $\mu$s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10): 2333–2341.

Chen, G.; Peng, P.; Li, G.; and Tian, Y. 2023. Training full spike neural networks via auxiliary accumulation pathway. *arXiv preprint arXiv:2301.11929*.

Delbrück, T.; Linares-Barranco, B.; Culurciello, E.; and Posch, C. 2010. Activity-driven, event-based vision sensors. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, 2426–2429. IEEE.

Dong, Y.; Li, Y.; Zhao, D.; Shen, G.; and Zeng, Y. 2024. Bullying10K: a large-scale neuromorphic dataset towards privacy-preserving bullying recognition. *Advances in Neural Information Processing Systems*, 36.

Fang, W.; Chen, Y.; Ding, J.; Yu, Z.; Masquelier, T.; Chen, D.; Huang, L.; Zhou, H.; Li, G.; and Tian, Y. 2023. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40): eadi1480.

Fang, W.; Yu, Z.; Chen, Y.; Huang, T.; Masquelier, T.; and Tian, Y. 2021a. Deep residual learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 34: 21056–21069.

Fang, W.; Yu, Z.; Chen, Y.; Masquelier, T.; Huang, T.; and Tian, Y. 2021b. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2661–2671.

Gerstner, W.; Kistler, W. M.; Naud, R.; and Paninski, L. 2014. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press.

Han, J.; Asano, Y.; Shi, B.; Zheng, Y.; and Sato, I. 2023. High-fidelity event-radiance recovery via transient event frequency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20616–20625.

Joo, H. K.; Vo, K.; Yamazaki, K.; and Le, N. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, 3230–3234. IEEE.

Kim, J.; Bae, J.; Park, G.; Zhang, D.; and Kim, Y. M. 2021. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2146–2156.

Li, H.; Liu, H.; Ji, X.; Li, G.; and Shi, L. 2017. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11: 309.

Li, W.; Mahadevan, V.; and Vasconcelos, N. 2013. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1): 18–32.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008. A 128×128 120 dB 15$\mu$ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2): 566–576.

Lin, Y.; Ding, W.; Qiang, S.; Deng, L.; and Li, G. 2021. Es-imagenet: A million event-stream classification dataset for spiking neural networks. *Frontiers in neuroscience*, 15: 726582.

Liu, Q.; Xing, D.; Tang, H.; Ma, D.; and Pan, G. 2021. Event-based Action Recognition Using Motion Information and Spiking Neural Networks. In *IJCAI*, 1743–1749.

Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, 2720–2727.

Luo, W.; Liu, W.; and Gao, S. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, 341–349.

Lv, H.; Zhou, C.; Cui, Z.; Xu, C.; Li, Y.; and Yang, J. 2021. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30: 4505–4515.

Miao, S.; Chen, G.; Ning, X.; Zi, Y.; Ren, K.; Bing, Z.; and Knoll, A. 2019. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in neurorobotics*, 13: 38.

Narayan, S.; Cholakkal, H.; Khan, F. S.; and Shao, L. 2019. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8679–8687.

Orchard, G.; Jayawant, A.; Cohen, G. K.; and Thakor, N. 2015. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9: 437.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance

deep learning library. *Advances in neural information processing systems*, 32.

Posch, C.; Matolin, D.; and Wohlgenannt, R. 2010. 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.

Ramachandra, B.; and Jones, M. 2020. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2569–2578.

Shen, S.; Zhao, D.; Shen, G.; and Zeng, Y. 2024. TIM: An Efficient Temporal Interaction Module for Spiking Transformer. *arXiv preprint arXiv:2401.11687*.

Sironi, A.; Brambilla, M.; Bourdis, N.; Lagorce, X.; and Benosman, R. 2018. HATS: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1731–1740.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4975–4986.

Wan, B.; Fang, Y.; Xia, X.; and Mei, J. 2020. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE international conference on multimedia and expo (ICME)*, 1–6. IEEE.

Wang, X.; Wu, Z.; Jiang, B.; Bao, Z.; Zhu, L.; Li, G.; Wang, Y.; and Tian, Y. 2024. Hardvs: Revisiting human activity recognition with dynamic vision sensors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5615–5623.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 322–339. Springer.

Zhang, J.; Dong, B.; Zhang, H.; Ding, J.; Heide, F.; Yin, B.; and Yang, X. 2022. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 8801–8810.

Zhang, X.; Yu, L.; Yang, W.; Liu, J.; and Xia, G.-S. 2023. Generalizing event-based motion deblurring in real-world scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10734–10744.

Zhou, C.; Yu, L.; Zhou, Z.; Ma, Z.; Zhang, H.; Zhou, H.; and Tian, Y. 2023. Spikingformer: Spike-driven residual learning for transformer-based spiking neural network. *arXiv preprint arXiv:2304.11954*.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3769–3777.

Zhou, J.; Zheng, X.; Lyu, Y.; and Wang, L. 2024. ExACT: Language-guided Conceptual Reasoning and Uncertainty Estimation for Event-based Action Recognition and More. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18633–18643.