frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 26 November 2024
DOI 10.3389/fncom.2024.1508297

OPEN ACCESS

EDITED BY
Cong Shi,
Chongqing University, China

REVIEWED BY
Man Yao,
Chinese Academy of Sciences (CAS), China
Peng Feng,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Hong Chen
✉ hongchen@tsinghua.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 09 October 2024
ACCEPTED 04 November 2024
PUBLISHED 26 November 2024

CITATION
Lin X, Liu M and Chen H (2024) Spike-HAR++: an energy-efficient and lightweight parallel spiking transformer for event-based human action recognition.
Front. Comput. Neurosci. 18:1508297.
doi: 10.3389/fncom.2024.1508297

Spike-HAR++: an energy-efficient and lightweight parallel spiking transformer for event-based human action recognition

Xinxu Lin[1,2,3†], Mingxuan Liu[4†] and Hong Chen[1,3*]

[1]School of Integrated Circuits, Tsinghua University, Beijing, China, [2]State Key Laboratory of Integrated Chips and Systems, Frontier Institute of Chip and System, Fudan University, Shanghai, China, [3]Greater Bay Area National Center of Technology Innovation, Research Institute of Tsinghua University in Shenzhen, Shenzhen, China, [4]School of Biomedical Engineering, Tsinghua University, Beijing, China

Event-based cameras are suitable for human action recognition (HAR) by providing movement perception with highly dynamic range, high temporal resolution, high power efficiency and low latency. Spike Neural Networks (SNNs) are naturally suited to deal with the asynchronous and sparse data from the event cameras due to their spike-based event-driven paradigm, with less power consumption compared to artificial neural networks. In this paper, we propose two end-to-end SNNs, namely Spike-HAR and Spike-HAR++, to introduce spiking transformer into event-based HAR. Spike-HAR includes two novel blocks: a spike attention branch, which enables model to focus on regions with high spike rates, reducing the impact of noise to improve the accuracy, and a parallel spike transformer block with simplified spiking self-attention mechanism, increasing computational efficiency. To better extract crucial information from high-level features, we modify the architecture of the spike attention branch and extend it in Spike-HAR to a higher dimension, proposing Spike-HAR++ to further enhance classification performance. Comprehensive experiments were conducted on four HAR datasets: SL-Animals-DVS, N-LSA64, DVS128 Gesture and DailyAction-DVS, to demonstrate the superior performance of our proposed model. Additionally, the proposed Spike-HAR and Spike-HAR++ require only 0.03 and 0.06 mJ, respectively, to process a sequence of event frames, with model sizes of only 0.7 and 1.8 M. This efficiency positions it as a promising new SNN baseline for the HAR community. Code is available at Spike-HAR++.

KEYWORDS

spiking neural network, human action recognition, transformer, attention branch, event-based vision

1 Introduction

Human action recognition (HAR) involves identifying and understanding human movements and has numerous applications in the real world (Sun et al., 2022). For instance, HAR can be employed in visual surveillance systems to detect hazardous activities and monitor human behavior, thereby ensuring safe operations (Lin et al., 2008). Additionally, HAR can facilitate sign language recognition (SLR). According to the latest data from the World Federation of the Deaf, there are 70 million deaf individuals worldwide using over 200 sign languages (Murray, 2018). However, learning sign language can be challenging and time-consuming, creating communication barriers for the deaf community (Hu L. et al., 2023).

To address this issue, HAR for sign language recognition has been extensively researched. Most of the works focused on using RGB or gray-scale videos as input for HAR (Wang et al., 2017; Kındıroglu et al., 2022; Vázquez-Enríquez et al., 2021; Mercanoglu Sincan and Keles, 2022; Shen et al., 2024; Wang F. et al., 2023), due to their popularity and easy access. However, the recognition results of RGB-based HAR methods are inevitably influenced by the motion blur inherent to RGB cameras and static background noise (Wang et al., 2019; Wang Y. et al., 2022).

As an emerging neuromorphic sensor, the event camera detects changes in brightness for each pixel independently, generating an event stream asynchronously and sparsely. The difference between RGB video frames [from LSA64 (Ronchetti et al., 2023)] and DVS event frames (from N-LSA64) is shown in Figure 1. The event camera features high temporal resolution, low latency, low power consumption, and a wide dynamic range (Su et al., 2022), which can effectively address issues related to motion blur and static background noise. That is, event cameras hold significant advantages in the field of HAR. The current state-of-the-art (SOTA) approaches for event-based HAR involve firstly designing event aggregation strategies converting the asynchronous output of the event camera into synfirst chronous visual frames, followed by processing using Artificial Neural Networks (ANNs) (Ghosh et al., 2019; Amir et al., 2017; Baldwin et al., 2022; Cannici et al., 2020; Innocenti et al., 2021; Sabater et al., 2022), which require considerable computational power, posing challenges for deployment on edge devices.

As third-generation neural networks, Spike Neural Networks (SNNs) are designed with biological plausibility, mimicking the dynamics of brain neurons to encode and transmit information in the form of spikes (Maass, 1997). Compared to ANNs, the event-driven nature of SNNs significantly reduces energy consumption when running on neuromorphic chips (Zhang et al., 2023, 2021). However, current SNN-based HAR tasks still face challenges of lack of datasets and low recognition accuracy (Shi et al., 2023).

In this paper, we propose two models, Spike-HAR and Spike-HAR++, to simultaneously reduce power consumption and enhance recognition accuracy in event-based HAR. Spike-HAR integrates a patch embedding (PE) block, parallel transformer blocks, a spike attention branch, and a classification head. To further improve performance, we modify the architecture and position of spike attention branch in Spike-HAR according to the Hu et al. (2024) and extend it to a higher dimension, proposing Spike-HAR++, which enables better extraction of crucial information from high-level features. As illustrated in Figure 2, experiments on the SL-Animals-DVS dataset (Vasudevan et al., 2022) demonstrate that both models significantly outperform other event-based HAR systems while maintaining lower levels of power consumption.

This paper is an extended version of our prior work (Lin et al., 2024) accepted by BMVC 2024. The main differences with the conference version are as follows: (1) besides the Spike-HAR based on the Parallel Spiking Transformer (referred to as Spike-SLR in the BMVC version), we newly propose Spike-HAR++, which is better suited for recognizing long-duration actions; (2) the application scope of the models are extended from sign language recognition to human action recognition, with comprehensive testing conducted on two additional datasets: DVS128 Gesture (Amir et al., 2017) and DailyAction-DVS (Liu et al., 2021), achieving SOTA performance; (3) a detailed overview about traditional ANN-based and SNN-based HAR methods, as well as the development of spiking transformers are discussed in the related work. To sum up, the main contributions of this paper are listed:

(1) We propose the Spike-HAR family, i.e., Spike-HAR and Spike-HAR++, which mainly consists of a powerful parallel spike transformer block. To the best of our knowledge, it is the first spiking transformer specifically designed for event-based HAR. To enhance the model's spatio-temporal attention to fine-grained action features while maintaining energy efficiency and a lightweight design, we employ a parallel spiking transformer. In this architecture, multi-layer perceptrons (MLPs) and simplified attention sub-modules (CB-S3A) operate in parallel to improve overall efficiency.

(2) We first introduce attention mask mechanisms into SNNs and incorporate a spike attention branch in our model to extract key regions from the input event streams. Additionally, we improve the attention operation for Spike-HAR++, utilizing high-dimensional features extracted through a patch embedding (PE) block to accommodate the recognition of long-duration actions. Experiments demonstrate that, although the parameter count and power consumption of Spike-HAR++ increase slightly, the accuracy of HAR improves significantly.

(3) Experimental results on the public datasets SL-Animals-DVS (Vasudevan et al., 2022), N-LSA64 (Ronchetti et al., 2023) [converted using the v2e (Hu et al., 2021) method], DVS128 Gesture (Amir et al., 2017), and DailyAction-DVS (Liu et al., 2021) show that the proposed Spike-HAR family effectively balances model size and recognition accuracy. Specifically, the proposed Spike-HAR and Spike-HAR++ require only 0.03 and 0.06 mJ, respectively, to process a sequence of event frames, with model size of just 0.7 and 1.8 M.

In the rest of the paper, Section 2 presents the related work on event-based HAR and spiking transformers. Section 3 begins with an overview of the overall architecture of Spike-HAR and Spike-HAR++, followed by a detailed description of each model component. Section 4 introduces four HAR benchmark datasets and evaluation metrics, along with rigorous ablation studies, visualizations, and performance evaluations of the proposed models. Finally, Section 5 concludes the paper.

2 Related work

2.1 Event-based human action recognition

Human action recognition aims to assign labels to various human behaviors and has wide applications in fields such as visual surveillance systems (Prati et al., 2019; Lin et al., 2008; Nasir et al., 2022), sign language recognition (Lin et al., 2024), autonomous navigation systems (Wang Q. et al., 2022), and video retrieval (Sahoo et al., 2020). Traditional HAR methods commonly use RGB or grayscale video as input due to their accessibility. However, HAR based on RGB modalities is not robust to illumination changes and is susceptible to motion artifacts (Sun et al., 2022). Additionally,

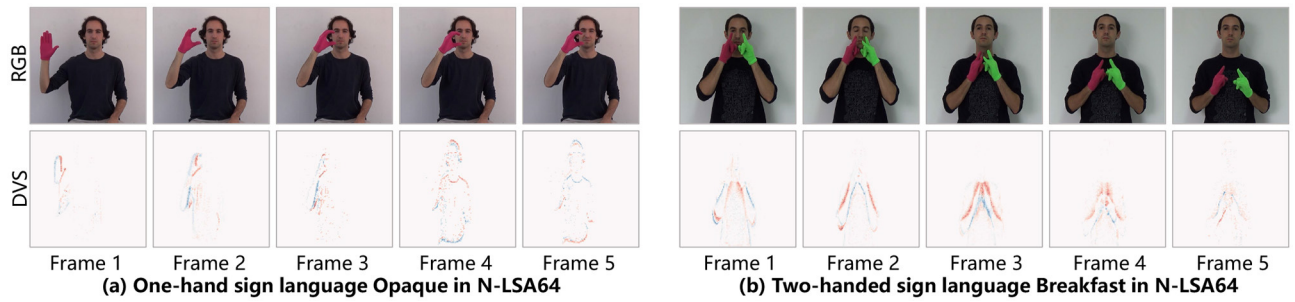

FIGURE 1
(a) Comparison of RGB video frames and DVS data frames for sign language Opaque (one-handed sign). **(b)** Comparison of RGB video frames and DVS data frames for sign language breakfast (two-handed sign).

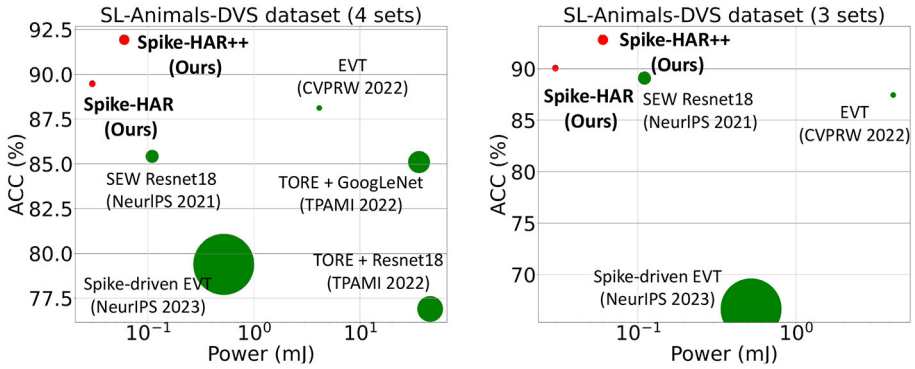

FIGURE 2
Accuracy vs. inference energy of different neural methods implemented in Intel Stratix 10 TX (Corporation, 2023) (for ANNs) or ROLLS (Qiao et al., 2015) (for SNNs). The size of the markers denotes the number of parameters.

the large data size of RGB videos results in high computational costs when modeling spatiotemporal context for HAR. To address above problem, alternative data forms for HAR have emerged, such as skeleton (Wang and Yan, 2023), depth (Sahoo et al., 2020), infrared sequences (Ding et al., 2022), point clouds (Yu et al., 2022), and event streams. This study focuses on event-based HAR, as event cameras offer high dynamic range, low latency, low power consumption, and eliminate motion blur, making them well-suited for HAR. Furthermore, the captured frames typically lack background information, which aids in action understanding.

The methods for event-based HAR can be primarily categorized into ANN-based and SNN-based (Gao et al., 2023). For ANN-based methods, representative studies mainly utilize 3D CNNs or transformers to learn features in both spatial and temporal domains, thereby aggregating information from adjacent frames. For example, Wang et al. (2024) presented a novel event stream-based action recognition model called EVMamba, which integrates a spatial plane multi-directional scanning mechanism with an innovative voxel temporal scanning mechanism to effectively extract spatio-temporal information from event streams. Acin et al. (2023) introduced VK-SITS, a new event data representation using the ResNet18 network, which outperformed other methods such as TORE (Baldwin et al., 2022) and SITS (Manderscheid et al., 2019). Additionally, Sabater et al. (2022) developed EVT, an efficient transformer model that leverages the sparsity of event data, achieving SOTA results on the SL-Animals-DVS dataset. They further improved EVT by employing a finer patch-based event data representation with richer spatio-temporal information, resulting in the introduction of the EVT+ model (Sabater et al., 2023). Gao et al. (2023) proposed the EV-ACT framework, which consists of an event voxel filtering module, a learnable multi-representation fusion module, an event-based slow-fast network, and an event-based spatio-temporal attention mechanism. This framework was tested on a new event-based HAR benchmark called THU$^{\text{E-ACT}}$-50 and its accompanying dataset, THU$^{\text{E-ACT}}$-50-CHL. Although ANN-based methods have achieved SOTA performance, they often involve high power consumption and a large number of model parameters due to the large data volume and significant information redundancy introduced by the temporal dimension, making them less suitable for edge applications in HAR. To address the problem, SNN-based methods have been proposed, leveraging their inherent temporal dynamics and energy efficiency. Specifically, Vasudevan et al. (2022) introduced the SL-Animals-DVS dataset and evaluated three types of SNNs, including SLAYER (Shrestha and Orchard, 2018), STBP (Wu et al., 2018), and DECOLLE (Kaiser et al., 2020), where the

test accuracy for all models remained below 75%. Liu et al. (2021) were the first to apply motion information in SNNs for event-based action recognition, surpassing existing SNN methods on three datasets, including DailyAction-DVS. Although SNNs can achieve energy-efficient recognition, they often yield suboptimal results.

2.2 Spiking transformers

ANN-based transformers have achieved success in fields such as vision and natural language processing (NLP) (Achiam et al., 2023; Han et al., 2022). However, the exploration of self-attention (SA) mechanisms based on SNNs remains limited, primarily because the multiplication operations inherent in vanilla self-attention (VSA) mechanism (Vaswani et al., 2017) are incompatible with SNNs. Recently, research has increasingly focused on developing the spiking transformer, aiming at eliminating multiplication operations in SA to reduce computational complexity. Zhou et al. (2022) were the first to introduce spiking transformer model, termed Spikformer, which utilizes spike-based Query, Key, and Value to model sparse visual features, thereby avoiding softmax computations. Subsequently, Yao et al. (2024b) introduced the Spike-driven Transformer, which enhances the spiking self-attention (SSA) mechanism in the Spikeformer. They proposed a Spike Driven Self-Attention (SDSA) that utilizes only masking and addition to implement the SA mechanism, reducing the computational complexity from $O(ND^2)$ to $O(ND)$. Wang Z. et al. (2023) introduced a novel Masked Spike Transformer (MST) framework, incorporating a Random Spike Masking (RSM) method, to further prune redundant spikes and reduce energy consumption without sacrificing performance. These exploration of spiking transformers enhance the learning capabilities of SNNs, enabling their application in various fields such as audio-visual classification, human pose tracking, and remote photoplethysmography (Guo et al., 2023; Zou et al., 2023; Liu et al., 2024). However, there is a lack of spiking transformers specifically designed for event-based HAR. We are the first to propose Spike-HAR (Lin et al., 2024), which is primarily composed of an energy-efficient parallel spiking transformer and has been tested on two DVS sign language datasets. Subsequently, SVFormer (Yu et al., 2024) was introduced as a direct training spiking transformer for efficient video action recognition, but it mainly focuses on RGB-based HAR. Wang X. et al. (2023) proposed a model called SSTFormer, which bridges SNNs and memory support transformers. However, SSTFormer is a hybrid SNN-ANN network requires both RGB frames and event streams to perform HAR. Therefore, dedicated spiking transformer models for event-based HAR still require further investigation and validation on larger-scale datasets.

3 Methodology

The proposed Spike-HAR and Spike-HAR++ apply the spiking transformer to HAR tasks. We utilize the SNNs algorithm provided in the SpikingJelly platform (Fang et al., 2023), employing the Leak Integrate and Fire (LIF) (Stein and Hodgkin, 1967) neural model for constructing the spiking neuron layers. LIF can be simply expressed by the following equation:

$$H[t] = V[t-1] + \frac{1}{\tau}(X[t] - (V[t-1] - V_{reset})) \tag{1}$$

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

$$V[t] = H[t](1 - S[t]) + V_{reset}S[t] \tag{3}$$

where t denotes the timestep, τ represents the membrane time constant, $X[t]$ donates the synaptic input current at time step t, and $H[t]$ is the neuron's membrane potential post charging and pre-spike, derived by integrating the input current. The spike occurrence at time t, denoted by $S[t]$, is determined by the Heaviside step function Θ, which outputs a spike (value of 1) when $H[t]$ surpasses the firing threshold V_{th}, indicating an action potential. $V[t]$ represents the membrane potential after spiking, which equals to $H[t]$ if no spike occurs and otherwise reset to V_{reset}.

3.1 Overall architecture

To lighten the models, we adopt less weight parameters and simpler model structures. The parameters of Spike-HAR and Spike-HAR++ is provided in Table 6, which are less than most models. In terms of model structure, Spike-HAR and Spike-HAR++ use a more simplified data preprocessing layer compared to the Spiking Transformer. And in both models we only use two MLP layers. Figure 3a illustrates the structure of Spike-HAR and Spike-HAR++, both of which consist of four main components: the patch embedding (PE) block, the parallel spike-driven transformer block, the spike attention branch, and the classification head. The PE block extracts spatio-temporal representations from the input DVS frames, while the CB-S3A module in the transformer and the spike firing rate map in the spike attention branch direct the model's focus toward key features. The final prediction head maps these features to possible sign language expressions.

Given a 2D DVS frames sequence $I_0 \in \mathbb{R}^{T_0 \times 2 \times H_0 \times W_0}$, where T_0, 2, H_0, W_0 represent the time step, initial number of channels, height and weight respectively. Firstly we randomly select continuous event frames with a time step of T ($T \leqslant T_0$) and crop each event frame spatially to obtain the preprocessed frames (PR), denoted as $I \in \mathbb{R}^{T \times 2 \times H \times W}$. The SNN-Based PE block, consisting of four 2D convolutional (Conv2D) layers, three batch normalization (BN) layers, three SNN layers and two max pooling (MP) layers, downsamples the input frames and partitioning them into spatio-temporal spike tokens $S_{PE} \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}}$, where D represents the number of channels. Before entering the data into the parallel Spike-driven Transformer block, we use membrane potential residual connection to avoid network degradation, adding S_{PE} and the output I_{PE} of the initial three convolutional layers and resulting the input S_0 of the same

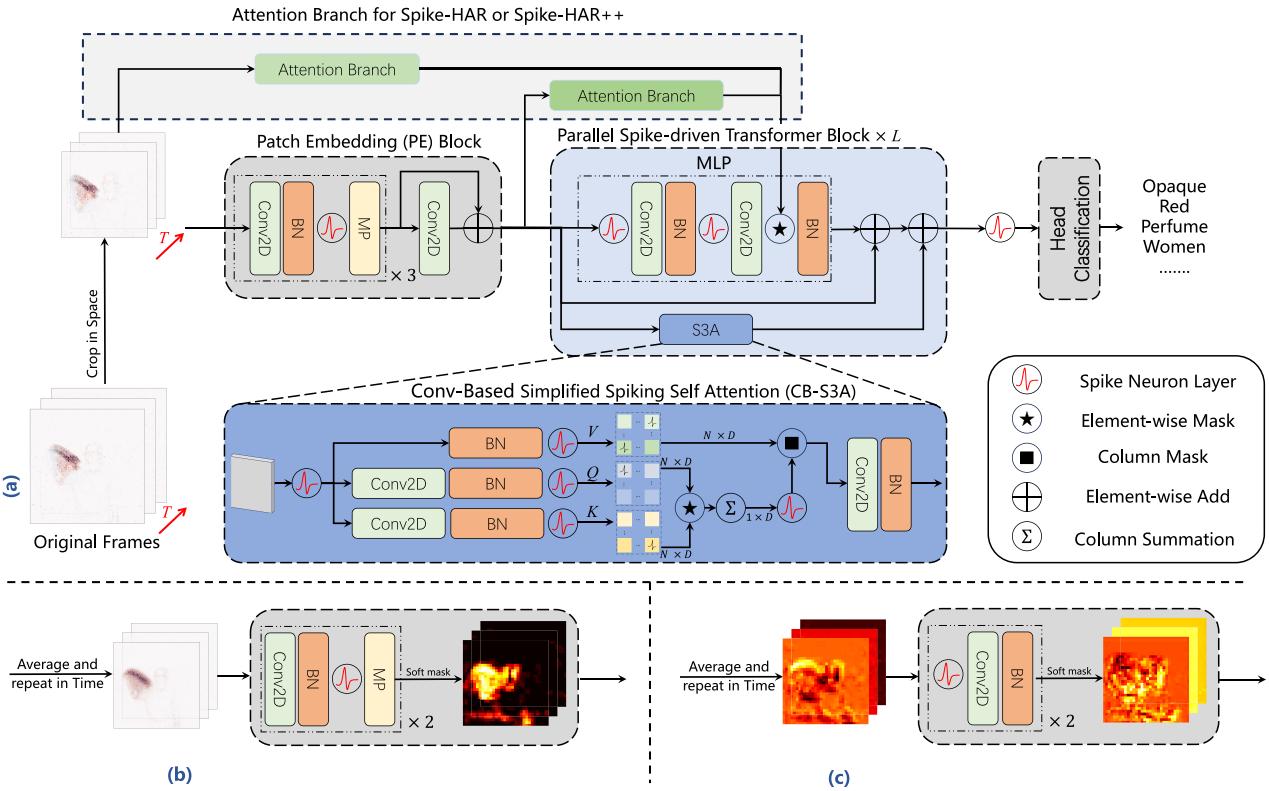

FIGURE 3
Framework of Spike-HAR and Spike-HAR++. We follow the network structure in Yao et al. (2024b). It consists of an SNN-based patch embedding (PE) block, several parallel spike-driven transformer blocks, a spike attention branch, and a SNN-based predictor head. **(a)** Architecture of Spike-HAR and Spike-HAR++. **(b)** Attention branch for Spike-HAR. **(c)** Attention branch for Spike-HAR++.

shape as S_{PE}. Therefore, the SNN-based PE block can be written as follows:

$$I = \mathrm{PR}(I_0) \quad I_0 \in \mathbb{R}^{T_0 \times 2 \times H_0 \times W_0}, I \in \mathbb{R}^{T \times 2 \times H \times W} \tag{4}$$

$$S_{PE} = \mathrm{PE}(I) \quad S_{PE} \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}} \tag{5}$$

$$S_0 = I_{PE} + S_{PE} \quad S_0 \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}} \tag{6}$$

Then, the spike sequence S_0 is passed to the parallel spike-driven transformer blocks, which consists of a conv-based simplified spiking self-attention (CB-S3A) block and a MLP block. As the main component in Spike-HAR and Spike-HAR++, CB-S3A, which just performs the convolution operation in spike-form Query (Q) and Key (K), offers an efficient method to model the local-global information of frames without softmax. In addition, the spike fire map generated by the spike attention branch performs mask operation on the data produced by the second convolution in the MLP block, which makes model more focus on local features. The outputs of the MLP and the CB-S3A blocks are summed together, and the sum is then added to the input S_0 again using membrane potential residual connection (RES). After L transformer blocks, the final output membrane potentials S_L is obtained. To obtain the pulse expression just consisting of 0 and 1, S_L then is passed to a spike neural layer ($\mathcal{SN}$), resulting in S_E. Finally, the S_E will be sent to a SNN-based classification head (SCH) to output the classification result Y. To summary, the output of CB-S3A, MLP and SCH can be written as follows:

$$S_l = \text{CB-S3A}(S_{l-1}) + \mathrm{MLP}(S_{l-1}) + S_{l-1} \tag{7}$$

$$S_l \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}}, l = 0...L$$

$$S_E = \mathcal{SN}(S_L) \quad S_E \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}} \tag{8}$$

$$Y = SCH(S_E) \tag{9}$$

3.2 Attention masks

DVS data can be influenced by noise from various sources, such as environmental background noise. As neural networks deepening, some noise may be amplified, causing the model to focus on irrelevant features. Inspired by Liu X. et al. (2020), we

insert attention blocks into our model to minimize the negative impact of background noise while allowing the model to focus on the target area and local features. In order to take note of the difference among different body parts, attention mask is applied to assign higher weights to pixels with stronger spike signals, while it is also the bridge between attention appearance and the backbone network. The difference between Spike-HAR and Spike-HAR++ lies in their implementation of attention branch. In Spike-HAR, the attention map is generated by directly averaging at the frame level, while Spike-HAR++ performs information extraction at a high-dimensional feature scale. The implementation methods of both are described in detail below.

3.2.1 Spike-HAR

As shown in Figure 3b, unlike the data processing operations performed in the PE block, we first perform a sum-average-repeat (SAR) operation on the data in the attention appearance. Specifically, we sum the event frames in the time dimension to combine multiple frames $I \in \mathbb{R}^{T\times 2\times H\times W}$ into a single frame $I_{SIN} \in \mathbb{R}^{1\times 2\times H\times W}$. Then, we divide the frame data by time step to obtain the average frame $I_{AVG} \in \mathbb{R}^{1\times 2\times \hat{H}\times \hat{W}}$ and replicate the I_{AVG} in the time dimension for T times as the input to the spike attention branch. The data $I_E \in \mathbb{R}^{T\times 2\times \hat{H}\times \hat{W}}$ undergoes two rounds of convolution and downsampling, followed by another SAR operation to obtain a spike fire rate map, which is then masked with the data in the MLP to facilitate communication between the branch and the backbone network as shown in Figure 3a.

3.2.2 Spike-HAR++

Directly summing event frames along the temporal dimension can efficiently aggregate critical spatial information at a low cost. However, it may fail during significantly prolonged actions. To address this issue, we utilize the spatio-temporal spike tokens $S_{PE} \in \mathbb{R}^{T\times D\times \frac{H}{4}\times \frac{W}{4}}$ extracted from the PE block and perform a SAR operation along the temporal dimension. These tokens are subsequently fed into a new spike attention branch, where they undergo two LIF-Conv-BN operations (shown in Figure 3c), followed by averaging along the temporal dimension to produce the attention mask. By leveraging the key features extracted by the PE block, the generated multi-channel mask is more representative. Experiments (Section 4) demonstrate that, although this adjustment increases power consumption by 0.03 mJ and model complexity as the convolution block must handle a larger number of feature channels, it significantly enhances HAR accuracy across various datasets.

3.3 Parallel spike-driven transformer

In the previous spiking Transformer architecture (Zhou et al., 2022; Yao et al., 2024b,a), the output U_{out} of the backbone network is transformed from the input U_{in} consisting of N tokens with dimension D using two consecutive sub-blocks (one SA and one MLP) with residual connections:

$$U_{\text{out}} = \alpha_{\text{FF}}\hat{U} + \beta_{\text{FF}}\text{MLP}(\mathcal{SN}(\hat{U})) \tag{10}$$

$$\hat{U} = \alpha_{\text{SA}}U_{\text{in}} + \beta_{\text{SA}}\text{SA}(\mathcal{SN}(U_{\text{in}})) \tag{11}$$

where scalar gain weights α_{FF}, β_{FF}, α_{SA}, β_{SA} fixed to 1 by default. In our work, to simplify the transformer block, we remove the residual connections in the MLP sub-blocks, obtaining the following output:

$$S_{\text{out}} = \alpha_{comb}S_{\text{in}} + \beta_{\text{FF}}\text{MLP}(\mathcal{SN}(S_{\text{in}})) + \beta_{\text{SA}}\text{SA}(\mathcal{SN}(S_{\text{in}})) \tag{12}$$

with skip gain $\alpha_{comb} = 1$, and residual gains $\beta_{\text{FF}} = \beta_{\text{SA}} = 1$ as default. In the submodule CB-S3A, we first input the spike signals S_0 into the spike neuron layer to obtain $S^{'}$. Then, we use 2D convolution operations to extract spatial information separately, resulting in Q and K. The acquisition of V does not involve convolution operations. After that, we use the spike neuron layer again to transform Q, K, and V into spike tensors Q_S, K_S, and V_S. And the subsequent masking calculation can be represented as follows:

$$\begin{aligned} MASK\left(Q_S, K_S, V_S\right) &= g\left(Q_S, K_S\right) \otimes V_S \\ &= \mathcal{SN}\left(SUM_C\left(Q_S \otimes K_S\right)\right) \otimes V_S \end{aligned} \tag{13}$$

where $\otimes$ denotes the Hadamard product, $g(\cdot)$ is used to compute the attention map, and SUM_C is used to calculate the sum of each column. The outputs of $g(\cdot)$ and SUM_C are row vectors of dimension D. Additionally, the Hadamard product between pulse tensors is equivalent to mask computation.

4 Experimental evaluation

4.1 Dataset

We evaluate our models on three public datasets, all generated by recording actions in real scenes. SL-Animals-DVS (Vasudevan et al., 2022) and DVS128 Gesture (Amir et al., 2017) were captured by a 128 × 128 pixel DVS128 camera, while DailyAction-DVS (Liu et al., 2021) was captured by a DAVIS346 camera with a spatial resolution of 346 × 260. Furthermore, we also tested our models using the N-LSA64 dataset which is transformed from LSA64 (Ronchetti et al., 2023) dataset using v2e (Hu et al., 2021) method.

4.1.1 SL-Animals-DVS

In the SL-Animals-DVS dataset 59 individuals were recorded separately, and each individual performed 19 signs in sequence. Due to the fact that the recording is conducted in 4 sessions at different locations under different lighting conditions, it can be further divided into SL-Animals-DVS-4sets, which includes four shooting environments, and SL-Animals-DVS-3sets, which includes three shooting environments.

4.1.2 DVS128 gesture

The DVS128 Gesture dataset comprises 1,342 recordings of 29 subjects performing 11 different actions (including one rejected class with random gestures) under three different lighting conditions.

4.1.3 DailyAction-DVS

The DailyAction-DVS dataset comprises 1,440 recordings of 15 subjects acting 12 different actions, including *bend*, *climb*, *fall down*, *get up*, *jump*, *lie down*, *carry box*, *run*, *sit down*, *stand up*, *walk* and *pick up*.The actions were captured under two lighting conditions including *natural light* and *LED light*.

4.1.4 N-LSA64

The N-LSA64 contains 3,200 DVS videos in which 10 non-expert subjects performed five repetitions of 64 different types of sign language. The symbols were selected from the most commonly used symbols in the LSA lexicon, including verbs and nouns. Depending on the number of hands performing the sign language, we further divide the data into N-LSA64-Right, which includes only right-hand movements, and N-LSA64-Both, which includes movements involving both hands.

We utilize a frame-based representation to preprocess an event stream (Fang et al., 2021b; Yao et al., 2021), transforming it into a sequence of event frames. Suppose the interval between two frames (i.e., temporal resolution) is dt and there are T frames (i.e., timesteps), the total length of the input event stream is $t_{total} = dt \times T$ milliseconds. After processing these frames with the proposed model, we can obtain a prediction.

4.2 Implementation details

We set the number of parallel spike-driven transformer block $L = 2$ in Spike-HAR and Spike-HAR++. In the DVS128 Gesture datasets, the sample length, time step, and learning rate is set as 6,000 ms, 20 and $1 \times e^{-3}$ respectively. In the SL-Animals-DVS and N-LSA64 datasets, the sample length, time step, and learning rate is set as 500 ms, 10 and $1 \times e^{-4}$ respectively. In the DailyAction-DVS dataset, the sample length, time step, and learning rate is set as 1,200 ms, 10 and $1 \times e^{-3}$ respectively. For the training and evaluation of frame-based methods, if the number of frames contained in each event frames is larger than the timesteps T, we linearly sample T of them. Otherwise, we pad them to the length of T with the zero-padding operation. Spike-HAR and Spike-HAR++ are optimized with AdamW (Loshchilov and Hutter, 2017) optimizer, in a single NVIDIA GeForce RTX 3090. We set the batch size to 32 and trained for 240 epochs using the one cycle learning rate policy (Smith and Topin, 2018). As for the data augmentation, we use spatial and temporal random crop and repeat each sample within the training batch twice with different augmentations. In addition, for the N-LSA64 dataset, we divided the data into training, validation, and test sets in the ratio of 6:2:2, and evaluated the classification accuracy on the test set.

4.3 Comparison to the state-of-the-art models

We compare the proposed Spike-HAR and Spike-HAR++ with several relevant action recognition methods, including SNN and ANN. And the results on four datasets are shown in Tables 1–4, respectively. We can find that our proposed models outperform existing action recognition methods, indicating that our proposed models have a stronger ability to extract action information from event data. Specifically, on the SL-Animals-DVS dataset, we compare our models with existing ANN models, SNN models and a hybrid neural network that includes both ANN and SNN components. Additionally, we replace the backbone network in EVT (Sabater et al., 2022) with the Spike-Driven Transformer block (Yao et al., 2024b) to obtain Spike-Evt and conduct model training for comparative analysis. Experimental results on SL-Animals-DVS are given in Table 1, from which we can see that the accuracy of Spike-HAR++ is 3.81 and 5.37% higher than that of EVT (Sabater et al., 2022) on the dataset SL-Animals-DVS-4sets and SL-Animals-DVS-3sets, respectively. And compared to the SNN method EventRPG + SEW Resnet18 (Sun et al., 2024), the Spike-HAR++ improves the accuracy by 0.35% on the dataset SL-Animals-DVS-4sets. On the SL-Animals-DVS-3sets, the EventRPG+SEW Resnet18 achieves a higher classification accuracy of 93.30% by using complex data augmentation strategies. In contrast, Spike-HAR++ reaches a similar accuracy of 92.82% with simple data augmentation (Section 4.2) and a more lightweight backbone (Section 4.5, Spike-HAR++ vs. SEW ResNet18). On the N-LSA64-Both and N-LSA64-Right datasets, for comparison with existing methods, we adopt the same sampling and training strategies to train the SOTA ANN model EVT (Sabater et al., 2022), the baseline SNN model STBP (Wu et al., 2018), and Spike-EVT, which is constructed by replacing the EVT backbone with a spiking transformer (Yao et al., 2024b). The test results, presented in Table 2, demonstrate that Spike-HAR++ increases accuracy by 1.72% compared to EVT on the N-LSA64-Both dataset and by 5.71% compared to the Spike-driven EVT on the N-LSA64-Right dataset. Furthermore, compared to the other models, Spike-HAR and Spike-HAR++ utilize the shortest sample length of 500 ms. For the DVS128 Gesture, as can be seen in Table 3, Spike-HAR and Spike-HAR++ get the classification accuracy of 98.26 and 97.92%, respectively, outperforming other ANN and SNN methods. Finally, as shown in Table 4, we compared our results on DailyAction-DVS with state-of-the-art SNN models. Spike-HAR++ achieved the best classification performance, reaching 98.47%, using the sample length of just 1,200 ms.

4.4 Ablation study

In this section, we analyze the impact of hyperparameters and the key components of Spike-HAR and Spike-HAR++. Experiments are conducted on the SL-Animals-DVS-4sets dataset. With a fixed total sample length of 500 ms, different time steps are set to investigate the impact of the number of input event frames and transformer blocks on the model results. As can be seen in Figure 4, with the number of time steps and the number of MLP Blocks increasing, the test accuracy of the model does not change significantly, but with the number of time steps increasing to be more than 20 or the number of MLP blocks decreasing to be 1, the test accuracy will have a significant decrease. Specifically, the highest accuracy of 89.47% for Spike-HAR and 91.93% for

TABLE 1 Classification accuracy in the SL-Animals-DVS dataset.

Model	Method	Time step	Sample length	SL-Animals-DVS	
				4 sets	3 sets
TORE + GoogLeNet (Baldwin et al., 2022)	ANN	\\	\	0.8510	\
TORE + ResNet18 (Baldwin et al., 2022)	ANN	\	\	0.7690	\
VoxelGrid + ResNet18 (Zhu et al., 2019)	ANN	\	\	0.8902	\
SITS + ResNet18 (Manderscheid et al., 2019)	ANN	\	\	0.7847	\
VK-SITS + ResNet18 (Acin et al., 2023)	ANN	\	\	0.7926	\
EVT (Sabater et al., 2022)	ANN	\	504 ms	0.8812	0.8745
SCTFA + 7-Layer Spiking CNN (Cai et al., 2024)	Hybrid	\	\	0.9004	\
SLAYER (Shrestha and Orchard, 2018)	SNN	300	1,500 ms	0.5430	0.6141
STBP (Wu et al., 2018)	SNN	50	1,500 ms	0.6497	0.7147
DECOLLE (Kaiser et al., 2020)	SNN	500	500 ms	0.6219	0.6219
SEW Resnet18 (Fang et al., 2021a)	SNN	16	\	0.8542	0.8909
EventDrop + SEW ResNet18 (Gu et al., 2021)	SNN	\	\	0.8633	0.8899
NDA + SEW ResNet18 (Li et al., 2022)	SNN	\	\	0.8777	0.8955
EventRPG + SEW ResNet18 (Sun et al., 2024)	SNN	\	\	**0.9159**	0.9330
Spike-Driven EVT (Yao et al., 2024b)	SNN	11	504 ms	0.7939	0.6667
Spike-HAR (Ours)	SNN	10	500 ms	0.8947	0.9006
Spike-HAR++ (Ours)	SNN	10	500 ms	0.9193	**0.9282**

Red and bold indicate the best and second best performance.

TABLE 2 Classification accuracy in the N-LSA64 dataset.

Model	Method	Time step	Sample length	N-LSA64	
				Both	Right
EVT (Sabater et al., 2022)	ANN	\	504 ms	0.8406	0.8214
STBP (Wu et al., 2018)	SNN	50	1,500 ms	0.5969	0.5786
Spike-driven EVT (Yao et al., 2024b)	SNN	11	504 ms	0.7266	0.8262
Spike-HAR (Ours)	SNN	10	500 ms	**0.8469**	**0.8690**
Spike-HAR++ (Ours)	SNN	10	500 ms	0.8578	0.8833

Red and bold indicate the best and second best performance.

Spike-HAR++ are achieved by setting the time step to 10 and the number of blocks to 2. On the other hand, the accuracy decreases to 81.72% for Spike-HAR and 87.72% for Spike-HAR++ when the time step is set to 25, and setting the number of blocks to 1 results in an accuracy of 84.65% for Spike-HAR and 90.88% for Spike-HAR++. In addition, The experimental results verify the parallel structure and the attention appearance used in proposed models. As shown in Table 5, using both parallel transformers and attention brunch simultaneously yields the best accuracy in Spike-HAR and Spike-HAR++.

4.5 Energy consumption analysis

We use the SL-Animals-DVS dataset to estimate the energy required for proposed models to classify a DVS sign language video. We first determine the number of operations [SOPs (Zhou et al., 2022) for the SNN module] needed to complete this task:

$$\mathrm{FLOPs}_{Conv2D} = \left(k_n\right)^2 \cdot h_n \cdot w_n \cdot c_{n-1} \cdot c_n \tag{14}$$

$$\mathrm{SOPs}_{Conv2D} = fr \cdot T_s \cdot \mathrm{FLOPs}_{Conv2D} \tag{15}$$

where k_n is the kernel size, (t_n, h_n, w_n) is the output feature map size, c_{n-1} and c_n are the input and output channel numbers, respectively. fr and T_s denote the spike fire rate and timesteps, respectively. The fr is defined as the proportion of non-zero elements within the spike tensor. Practically, we set T_s to 10. Once SOPs for the SNN module are determined, we can further obtain the final energy consumption E by multiplying the SOPs with the platform's energy:

$$E_{\mathrm{SOPs}} = E_{\mathrm{AC}} \times \mathrm{SOPs} \tag{16}$$

TABLE 3 Classification accuracy in the DVS128 Gesture dataset.

Model	Method	Time step	Sample length	DVS128 Gesture
12 layers CNN (Amir et al., 2017)	ANN	120	120 ms	0.9260
Identify + Resnet34 (He et al., 2016)	ANN	\	\	0.9549
NDA + Resnet34 (Li et al., 2022)	ANN	\	\	0.9722
EventMix + Resnet34 (Shen et al., 2023)	ANN	\	\	0.9180
ShapeAug + Resnet34 (Bendig et al., 2024)	ANN	\	\	0.9170
EventDrop + Resnet34 (Gu et al., 2021)	ANN	\	\	0.9618
PLIF-SNN (Fang et al., 2021b)	SNN	20	6,000 ms	0.9760
Res-SNN-18 (Yao et al., 2021)	SNN	16	6,000 ms	0.9790
ASA-SNN (Yao et al., 2023)	SNN	20	6,000 ms	0.9770
Identify + SEW Resnet18 (Fang et al., 2021a)	SNN	\	\	0.9433
Eventmix + SEW Resnet18 (Shen et al., 2023)	SNN	\	\	0.9675
EventRPG + SEW Resnet18 (Sun et al., 2024)	SNN	\	\	0.9653
Identify + CSNN (Xu et al., 2018)	SNN	\	\	0.9375
NDA + CSNN (Li et al., 2022)	SNN	\	\	0.9583
EventAugmentation + CSNN (Gu et al., 2024)	SNN	\	\	0.9625
EventDrop + CSNN (Gu et al., 2021)	SNN	\	\	0.9444
Spike-HAR (Ours)	SNN	20	6,000 ms	0.9826
Spike-HAR++ (Ours)	SNN	20	6,000 ms	**0.9792**

Red and bold indicate the best and second best performance.

TABLE 4 Classification accuracy in the DailyAction-DVS dataset.

Model	Method	Time step	Sample length	DailyAction-DVS
Gabor-Tempotron SNN (Xiao et al., 2019)	SNN	\	\	0.6830
HMAX-SNN (Liu Q. et al., 2020)	SNN	\	\	0.7690
Motion-SNN (Liu et al., 2021)	SNN	\	\	0.9030
PLIF-SNN (Fang et al., 2021b)	SNN	36	4,320 ms	0.9250
ASA-SNN (Yao et al., 2023)	SNN	36	4,320 ms	0.9460
EHTI & MDTS-Tempotron SNN (Ding et al., 2024)	SNN	\	\	0.9608
Spike-HAR (Ours)	SNN	10	1,200 ms	**0.9826**
Spike-HAR++ (Ours)	SNN	10	1,200 ms	0.9847

Red and bold indicate the best and second best performance.

We use the same energy efficiency calculation scheme proposed by Hu Y. et al. (2023). The energy consumption is 12.5 pJ for each floating-point operation (FLOP) and is 77 fJ for each synaptic operation (SOP). As shown in Table 6, the Spike-HAR processes DVS frame data with a spatial size of 96 × 96 and a time step of 10 with only 0.03 mJ of power consumption. This represents a 99.27% energy reduction compared to EVT and is substantially lower than that of other baseline models. Furthermore, although Spike-HAR++ has a higher power consumption compared to Spike-HAR (0.06 vs. 0.03 mJ), it is still lower than that of other models and achieves higher performance than Spike-HAR across the SL-Animals-DVS, N-LSA64, DVS128 Gesture, and DailyAction-DVS datasets.

5 Conclusion

In this paper, we proprse an energy-efficient and lightweight Spike-HAR family for event-based human action recognition, to adaptively emphasize on local spatial features as well as temporal features. Spike-HAR and Spike-HAR++ surpass existing methods in accuracy on the SL-Animals-DVS, N-LSA64, DVS128 Gesture, and DailyAction-DVS datasets. Furthermore, Spike-HAR and Spike-HAR++ require only 0.03 and 0.06 mJ to recognize a single action event stream, reducing the power consumption of 99.27 and 98.55% compared to the Evt, respectively. It demonstrates the applicability of spiking transformers for human action recognition and their potential application in human-machine interaction and edge HAR devices. In the future, it is promising to develop

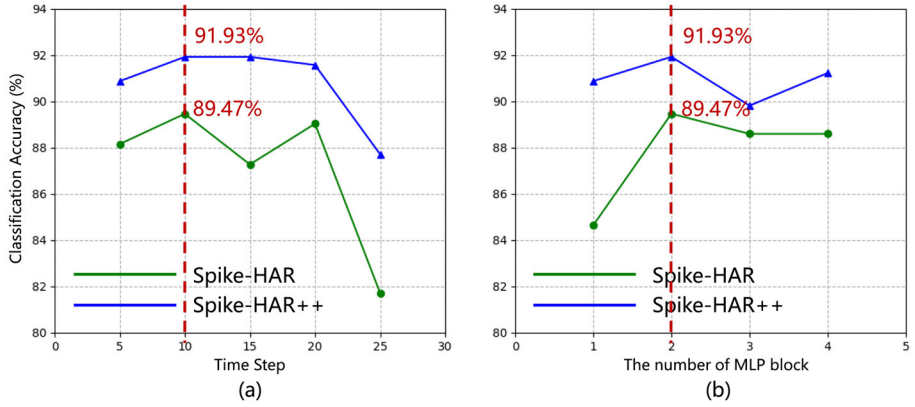

FIGURE 4
(a) Accuracy of Spike-HAR (green line) and Spike-HAR++ (blue line) at varying timesteps T (with 2 MLP blocks); **(b)** Accuracy of Spike-HAR (green line) and Spike-HAR++ (blue line) with different numbers of MLP blocks (with $T = 10$).

TABLE 5 Accuracy of Spike-HAR and Spike-HAR++ for different architecture on SL-animals-DVS-4sets.

Models	Attention brunch	Serial transformer block	Parallel transformer block	Accuracy
Spike-HAR		✓		0.8640
	✓	✓		0.8465
			✓	0.8421
	✓		✓	0.8947
Spike-HAR++		✓		0.8640
	✓	✓		0.9088
			✓	0.8421
	✓		✓	0.9193

TABLE 6 Computational complexity comparisons of SLR methods.

Model	Method	#Params.	FLOPs/SOPs	Power/mJ
TORE + ResNet18 (Baldwin et al., 2022)	ANN	11.69 M	3.66 G	45.75
TORE + GoogLeNet (Baldwin et al., 2022)	ANN	8.46 M	2.88 G	36.00
EVT (Sabater et al., 2022)	ANN	0.50 M	0.33 G	4.13
Spike-driven EVT (Yao et al., 2024b)	SNN	66.34 M	6.77 G	0.52
SEW Resnet18 (Fang et al., 2021a)	SNN	2.92 M	1.41 G	0.11
Spike-HAR (Ours)	SNN	0.70 M	0.44 G	0.03
Spike-HAR++ (Ours)	SNN	1.80 M	0.74 G	0.06

a more complex large-scale event-based HAR benchmark to further evaluate the performance of the Spike-HAR family in practical applications.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: SL-Animals-DVS (http://www2.imse-cnm.csic.es/neuromorphs/index.php/SL-ANIMALS-DVS-Database); LSA64 (https://facundoq.github.io/datasets/lsa64/); DVS128 Gesture (https://ibm.ent.box.com/s/3hiq58ww1pbbjrinh367ykfdf60xsfm8/folder/50167556794); DailyAction-DVS (https://github.com/qianhuiliu/SNN-action-recognition).

Ethics statement

Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article because the photographs appearing in the manuscript are sourced from publicly available datasets.

Author contributions

XL: Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft. ML: Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft. HC: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China under Grant 92164110 and 62334014.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. During the preparation of this work the author(s) used GPT-4 in order to polish the content. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Publisher's note

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. *arXiv* [Preprint]. arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774

Acin, L., Jacob, P., Simon-Chane, C., and Histace, A. (2023). "VK-sits: a robust time-surface for fast event-based recognition," in *2023 Twelfth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (Paris: IEEE), 1–6. doi: 10.1109/IPTA59101.2023.10320049

Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, 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* (Honolulu, HI: IEEE), 7243–7252. doi: 10.1109/CVPR.2017.781

Baldwin, R. W., Liu, R., Almatrafi, M., Asari, V., and Hirakawa, K. (2022). Time-ordered recent event (TORE) volumes for event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 2519–2532. doi: 10.1109/TPAMI.2022.3172212

Bendig, K., Schuster, R., and Stricker, D. (2024). Shapeaug: occlusion augmentation for event camera data. *arXiv* [Preprint]. arXiv:2401.02274. doi: 10.48550/arXiv.2401.02274

Cai, W., Sun, H., Liu, R., Cui, Y., Wang, J., Xia, Y., et al. (2024). A spatial-channel-temporal-fused attention for spiking neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 35, 14315–14329. doi: 10.1109/TNNLS.2023.3278265

Cannici, M., Ciccone, M., Romanoni, A., and Matteucci, M. (2020). "A differentiable recurrent surface for asynchronous event-based data," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX* (Berlin: Springer-Verlag), 136–152. doi: 10.1007/978-3-030-58565-5_9

Corporation, I. (2023). *Intel stratix 10 tx device overview*. Available at: https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/hb/stratix-10/s10_tx_overview.pdf (accessed December 10, 2023).

Ding, H., Jiang, J., and Yan, R. (2024). "A time-surface enhancement model for event-based spatiotemporal feature extraction," in *2024 International Joint Conference on Neural Networks (IJCNN)* (Yokohama IEEE), 1–7. doi: 10.1109/IJCNN60899.2024.10650047

Ding, M., Ding, Y., Wei, L., Xu, Y., and Cao, Y. (2022). Individual surveillance around parked aircraft at nighttime: thermal infrared vision-based human action recognition. *IEEE Trans. Syst. Man Cybern. Syst.* 53, 1084–1094. doi: 10.1109/TSMC.2022.3192017

Fang, W., Chen, Y., Ding, J., Yu, Z., Masquelier, T., Chen, D., et al. (2023). Spikingjelly: an open-source machine learning infrastructure platform for spike-based intelligence. *Sci. Adv.* 9:eadi1480. doi: 10.1126/sciadv.adi1480

Fang, W., Yu, Z., Chen, Y., Huang, T., Masquelier, T., Tian, Y., et al. (2021a). Deep residual learning in spiking neural networks. *Adv. Neural Inf. Process. Syst.* 34, 21056–21069. doi: 10.48550/arXiv.2102.04159

Fang, W., Yu, Z., Chen, Y., Masquelier, T., Huang, T., Tian, Y., et al. (2021b). "Incorporating learnable membrane time constant to enhance learning of spiking neural networks," in *Proceedings of the IEEE/CVF international conference on computer vision* (Montreal, QC: IEEE), 2661–2671. doi: 10.1109/ICCV48922.2021.00266

Gao, Y., Lu, J., Li, S., Ma, N., Du, S., Li, Y., et al. (2023). Action recognition and benchmark using event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 14081–14097. doi: 10.1109/TPAMI.2023.3300741

Ghosh, R., Gupta, A., Nakagawa, A., Soares, A., and Thakor, N. (2019). Spatiotemporal filtering for event-based action recognition. *arXiv* [Preprint] arXiv:1903.07067. doi: 10.48550/arXiv.1903.07067

Gu, F., Dou, J., Li, M., Long, X., Guo, S., Chen, C., et al. (2024). Eventaugment: learning augmentation policies from asynchronous event-based data. *IEEE Trans. Cogn. Dev. Syst.* 16, 1521–1532. doi: 10.1109/TCDS.2024.3380907

Gu, F., Sng, W., Hu, X., and Yu, F. (2021). Eventdrop: data augmentation for event-based learning. *arXiv* [Preprint]. arXiv:2106.05836. doi: 10.48550/arXiv.2106.05836

Guo, L., Gao, Z., Qu, J., Zheng, S., Jiang, R., Lu, Y., et al. (2023). Transformer-based spiking neural networks for multimodal audio-visual classification. *IEEE Trans. Cogn. Dev. Syst.* 16, 1077–1086. doi: 10.1109/TCDS.2023.3327081

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2022). A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 87–110. doi: 10.1109/TPAMI.2022.3152247

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Hu, L., Gao, L., Liu, Z., and Feng, W. (2023). "Continuous sign language recognition with correlation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC: IEEE), 2529–2539. doi: 10.1109/CVPR52729.2023.00249

Hu, Y., Deng, L., Wu, Y., Yao, M., and Li, G. (2024). Advancing spiking neural networks toward deep residual learning. *IEEE Trans. Neural Netw. Learn. Syst.* 1–15. doi: 10.1109/TNNLS.2024.3355393

Hu, Y., Liu, S.-C., and Delbruck, T. (2021). "v2e: From video frames to realistic dvs events," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Nashville, TN: IEEE), 1312–1321. doi: 10.1109/CVPRW53098.2021.00144

Hu, Y., Tang, H., and Pan, G. (2023). Spiking deep residual networks. *IEEE Trans. Neural Netw. Learn. Syst*. 34, 5200–5205. doi: 10.1109/TNNLS.2021.3119238

Innocenti, S. U., Becattini, F., Pernici, F., and Del Bimbo, A. (2021). "Temporal binary representation for event-based action recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)* (Milan: IEEE), 10426–10432. doi: 10.1109/ICPR48806.2021.9412991

Kaiser, J., Mostafa, H., and Neftci, E. (2020). Synaptic plasticity dynamics for deep continuous local learning (decolle). *Front. Neurosci*. 14:515306. doi: 10.3389/fnins.2020.00424

Kındıroglu, A., Özdemir, O., Akarun, L. (2022). Aligning accumulative representations for sign language recognition. *Mach. Vis. Appl*. 34, 1–18. doi: 10.1007/s00138-022-01367-x

Li, Y., Kim, Y., Park, H., Geller, T., and Panda, P. (2022). "Neuromorphic data augmentation for training spiking neural networks," in *European Conference on Computer Vision* (Cham: Springer), 631–649. doi: 10.1007/978-3-031-20071-7_37

Lin, W., Sun, M.-T., Poovandran, R., and Zhang, Z. (2008). "Human activity recognition for video surveillance," in *2008 IEEE international symposium on circuits and systems (ISCAS)* (Seattle, WA: IEEE), 2737–2740. doi: 10.1109/ISCAS.2008.4542023

Lin, X., Liu, M., Liu, K., and Chen, H. (2024). "Spike-slr: an energy-efficient parallel spiking transformer for event-based sign language recognition," in *BMVC 2024 - 2024 The British Machine Vision Conference (BMVC)* (Glasgow).

Liu, M., Tang, J., Li, H., Qi, J., Li, S., Wang, K., et al. (2024). Spiking-physformer: camera-based remote photoplethysmography with parallel spike-driven transformer. *arXiv* [Preprint]. arXiv:2402:04798. doi: 10.48550/arXiv.2402:04798

Liu, Q., Ruan, H., Xing, D., Tang, H., and Pan, G. (2020). Effective aer object classification using segmented probability-maximization learning in spiking neural networks. *Proc. AAAI Conf. Artif. Intell*. 34, 1308–1315. doi: 10.1609/aaai.v34i02.5486

Liu, Q., Xing, D., Tang, H., Ma, D., and Pan, G. (2021). "Event-based action recognition using motion information and spiking neural networks," in *International Joint Conferences on Artificial Intelligence Organization, Virtual conference (IJCAI)*, 1743–1749. doi: 10.24963/ijcai.2021/240

Liu, X., Fromm, J., Patel, S., and McDuff, D. (2020). "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Advances in Neural Information Processing Systems, Vol. 33*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Red Hook, NY: Curran Associates, Inc), 19400–19411.

Loshchilov, I., and Hutter, F. (2017). "Decoupled weight decay regularization," in *International Conference on Learning Representations* (Toulon).

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Netw*. 10, 1659–1671. doi: 10.1016/S0893-6080(97)00011-7

Manderscheid, J., Sironi, A., Bourdis, N., Migliore, D., and Lepetit, V. (2019). "Speed invariant time surface for learning to detect corner points with event-based cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 10245–10254. doi: 10.1109/CVPR.2019.01049

Mercanoglu Sincan, O., and Keles, H. Y. (2022). Using motion history images with 3D convolutional networks in isolated sign language recognition. *IEEE Access* 10, 18608–18618. doi: 10.1109/ACCESS.2022.3151362

Murray, J. (2018). *World federation of the deaf*. Available at: http://wfdeaf.org/our-work/ (accessed May 08, 2024).

Nasir, I. M., Raza, M., Shah, J. H., Wang, S.-H., Tariq, U., Khan, M. A., et al. (2022). Harednet: a deep learning based architecture for autonomous video surveillance by recognizing human actions. *Comput. Electr. Eng*. 99:107805. doi: 10.1016/j.compeleceng.2022.107805

Prati, A., Shan, C., and Wang, K. I.-K. (2019). Sensors, vision and networks: from video surveillance to activity recognition and health monitoring. *J. Ambient Intell. Smart Environ*. 11, 5–22. doi: 10.3233/AIS-180510

Qiao, N., Mostafa, H., Corradi, F., Osswald, M., Stefanini, F., Sumislawska, D., et al. (2015). A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses. *Front. Neurosci*. 9:123487. doi: 10.3389/fnins.2015.00141

Ronchetti, F., Quiroga, F. M., Estrebou, C., Lanzarini, L., and Rosete, A. (2023). Lsa64: an argentinian sign language dataset. *arXiv* [Preprint]. arXiv:2310.17429. doi: 10.48550/arXiv.2310.17429

Sabater, A., Montesano, L., and Murillo, A. C. (2022). "Event transformer. a sparse-aware solution for efficient event data processing," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (New Orleans, LA: IEEE), 2676–2685. doi: 10.1109/CVPRW56347.2022.00301

Sabater, A., Montesano, L., and Murillo, A. C. (2023). Event transformer+. a multi-purpose solution for efficient event data processing. *IEEE Trans. Pattern Anal. Mach. Intell*. 45, 16013–16020. doi: 10.1109/TPAMI.2023.3311336

Sahoo, S. P., Ari, S., Mahapatra, K., and Mohanty, S. P. (2020). Har-depth: a novel framework for human action recognition using sequential learning and depth estimated history images. *IEEE Trans. Emerg. Top. Comput. Intell*. 5, 813–825. doi: 10.1109/TETCI.2020.3014367

Shen, G., Zhao, D., and Zeng, Y. (2023). Eventmix: an efficient data augmentation strategy for event-based learning. *Inf. Sci*. 644:119170. doi: 10.1016/j.ins.2023.119170

Shen, X., Zheng, Z., and Yang, Y. (2024). Stepnet: apatial-temporal part-aware network for isolated sign language recognition. *ACM Trans. Multimedia Comput. Commun. Appl*. 20:226. doi: 10.1145/3656046

Shi, Q., Ye, Z., Wang, J., and Zhang, Y. (2023). Qisampling: an effective sampling strategy for event-based sign language recognition. *IEEE Signal Process. Lett*. 30, 768–772. doi: 10.1109/LSP.2023.3289111

Shrestha, S. B., and Orchard, G. (2018). Slayer: spike layer error reassignment in time. *Adv. Neural Inf. Process. Syst*. 31, 1412–1421.

Smith, L. N., and Topin, N. (2018). "Super-convergence: very fast training of neural networks using large learning rates," in *Defense + Commercial Sensing* (Baltimore, MD). doi: 10.1117/12.2520589

Stein, R., and Hodgkin, A. L. (1967). The frequency of nerve action potentials generated by applied currents. *Proc. R. Soc. Lond. B. Biol. Sci*. 167, 64–86. doi: 10.1098/rspb.1967.0013

Su, L., Yang, F., Wang, X., Guo, C., Tong, L., Hu, Q., et al. (2022). A survey of robot perception and control based on event camera. *Acta Autom. Sin*, 48, 1869–1889.

Sun, M., Zhang, D., Ge, Z., Wang, J., Li, J., Fang, Z., et al. (2024). Eventrpg: event data augmentation with relevance propagation guidance. *arXiv* [Preprint]. arXiv:2403.09274. doi: 10.48550/arXiv.2403.09274.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J., et al. (2022). Human action recognition from various data modalities: a review. *IEEE Trans. Pattern Anal. Mach. Intell*. 45, 3200–3225. doi: 10.1109/TPAMI.2022.3183112

Vasudevan, A., Negri, P., Di Ielsi, C., Linares-Barranco, B., and Serrano-Gotarredona, T. (2022). Sl-animals-dvs: event-driven sign language animals dataset. *Pattern Anal. Applic*. 25, 505–520. doi: 10.1007/s10044-021-01011-w

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Neural Information Processing Systems* (Long Beach, CA).

Vázquez-Enríquez, M., Alba-Castro, J. L., Docío-Fernández, L., and Rodríguez-Banga, E. (2021). "Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Nashville, TN: IEEE), 3457–3466. doi: 10.1109/CVPRW53098.2021.00385

Wang, C., and Yan, J. (2023). A comprehensive survey of RGB-based and skeleton-based human action recognition. *IEEE Access* 11, 53880–53898. doi: 10.1109/ACCESS.2023.3282311

Wang, F., Zhang, L., Yan, H., and Han, S. (2023). TIM-SLR: a lightweight network for video isolated sign language recognition. *Neural Comput. Appl*. 35, 22265–22280. doi: 10.1007/s00521-023-08873-7

Wang, P., Li, W., Gao, Z., Zhang, Y., Tang, C., Ogunbona, P., et al. (2017). "Scene flow to action map: a new representation for rgb-d based action recognition with convolutional neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI), 416–425. doi: 10.1109/CVPR.2017.52

Wang, Q., Luo, H., Wang, J., Sun, L., Ma, Z., Zhang, C., et al. (2022). Recent advances in pedestrian navigation activity recognition: a review. *IEEE Sens. J*. 22, 7499–7518. doi: 10.1109/JSEN.2022.3153610

Wang, Q., Zhang, Y., Yuan, J., and Lu, Y. (2019). "Space-time event clouds for gesture recognition: from RGB cameras to event cameras," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Waikoloa, HI: IEEE), 1826–1835. doi: 10.1109/WACV.2019.00199

Wang, X., Wang, S., Shao, P., Jiang, B., Zhu, L., Tian, Y., et al. (2024). Event stream based human action recognition: a high-definition benchmark dataset and algorithms. *arXiv* [Preprint]. arXiv:2408.09764. doi: 10.48550/arXiv.2408.09764

Wang, X., Wu, Z., Rong, Y., Zhu, L., Jiang, B., Tang, J., et al. (2023). Sstformer: bridging spiking neural network and memory support transformer for frame-event based recognition. *arXiv* [Preprint]. arXiv:2308.04369. doi: 10.48550/arXiv.2308.04369

Wang, Y., Zhang, X., Shen, Y., Du, B., Zhao, G., Cui, L., et al. (2022). Event-stream representation for human gaits identification using deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell*. 44, 3436–3449. doi: 10.1109/TPAMI.2021.3054886

Wang, Z., Fang, Y., Cao, J., Zhang, Q., Wang, Z., Xu, R., et al. (2023). "Masked spiking transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Paris: IEEE), 1761–1771. doi: 10.1109/ICCV51070.2023.00169

Wu, Y., Deng, L., Li, G., and Shi, L. (2018). Spatio-temporal backpropagation for training high-performance spiking neural networks. *Front. Neurosci*. 12:323875. doi: 10.3389/fnins.2018.00331

Xiao, R., Tang, H., Ma, Y., Yan, R., and Orchard, G. (2019). An event-driven categorization model for aer image sensors using multispike encoding and learning. *IEEE Trans. Neural Netw. Learn. Syst*. 31, 3649–3657. doi: 10.1109/TNNLS.2019.2945630

Xu, Q., Qi, Y., Yu, H., Shen, J., Tang, H., Pan, G., et al. (2018). "CSNN: an augmented spiking based framework with perceptron-inception," in *IJCAI, Vol. 1646* (Stockholm). doi: 10.24963/ijcai.2018/228

Yao, M., Gao, H., Zhao, G., Wang, D., Lin, Y., Yang, Z., et al. (2021). "Temporal-wise attention spiking neural networks for event streams classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 10221–10230. doi: 10.1109/ICCV48922.2021.01006

Yao, M., Hu, J., Hu, T., Xu, Y., Zhou, Z., Tian, Y., et al. (2024a). Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. *arXiv* [Preprint]. arXiv:2404.03663. doi: 10.48550/arXiv.2404.03663

Yao, M., Hu, J., Zhao, G., Wang, Y., Zhang, Z., Xu, B., et al. (2023). "Inherent redundancy in spiking neural networks," in *Proceedings of the IEEE/CVF international conference on computer vision* (Paris: IEEE), 16924–16934. doi: 10.1109/ICCV51070.2023.01552

Yao, M., Hu, J., Zhou, Z., Yuan, L., Tian, Y., Xu, B., et al. (2024b). Spike-driven transformer. *Adv. Neural Inf. Process. Syst.* 36, 64043–64058.

Yu, L., Huang, L., Zhou, C., Zhang, H., Ma, Z., Zhou, H., et al. (2024). Svformer: a direct training spiking transformer for efficient video action recognition. *arXiv* [Preprint]. arXiv:2406.15034. doi: 10.48550/arXiv.2406.15034

Yu, Z., Taha, A., Taylor, W., Zahid, A., Rajab, K., Heidari, H., et al. (2022). A radar-based human activity recognition using a novel 3-d point cloud classifier. *IEEE Sens. J.* 22, 18218–18227. doi: 10.1109/JSEN.2022.3198395

Zhang, J., Huo, D., Zhang, J., Qian, C., Liu, Q., Pan, L., et al. (2023). "22.6 anp-i: a 28nm 1.5 pj/sop asynchronous spiking neural network processor enabling sub-o. 1 μj/sample on-chip learning for edge-ai applications," *2023 IEEE International Solid-State Circuits Conference (ISSCC)* (San Francisco, CA: IEEE), 21–23. doi: 10.1109/ISSCC42615.2023.10067650

Zhang, J., Liang, M., Wei, J., Wei, S., and Chen, H. (2021). "A 28nm configurable asynchronous snn accelerator with energy-efficient learning," in *2021 27th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC)* (Beijing: IEEE), 34–39. doi: 10.1109/ASYNC48570.2021.00013

Zhou, Z., Zhu, Y., He, C., Wang, Y., Yan, S., Tian, Y., et al. (2022). Spikformer: when spiking neural network meets transformer. *arXiv* [Preprint]. arXiv:2209.15425. doi: 10.48550/arXiv.2209.15425

Zhu, A. Z., Yuan, L., Chaney, K., and Daniilidis, K. (2019). "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 989–997. doi: 10.1109/CVPR.2019.00108

Zou, S., Mu, Y., Zuo, X., Wang, S., and Li, C. (2023). Event-based human pose tracking by spiking spatiotemporal transformer. *arXiv* [Preprint]. arXiv:2303.09681. doi: 10.48550/arXiv.2303.09681