Spike-SLR: An Energy-efficient Parallel Spiking Transformer for Event-based Sign Language Recognition

Xinxu Lin *[1]
linxinxulx@foxmail.com
Mingxuan Liu *[2]
https://arktis2022.github.io/
Kezhuo Liu[1]
lkz22@mails.tsinghua.edu.cn
Hong Chen✉[1]
hongchen@tsinghua.edu.cn

[1] School of Integrated Circuits
Tsinghua University
Beijing, China

[2] School of Biomedical Engineering
Tsinghua University
Beijing, China

Abstract

Event-based cameras are suitable for sign language recognition (SLR) 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 introduce spiking transformer into event-based SLR by proposing a model named Spike-SLR, which includes two novel blocks: a spike soft-attention block, 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. On SL-Animals-DVS-4sets and SL-Animals-DVS-3sets, Spike-SLR achieves the accuracy of 89.47% and 90.06%, outperforming the state-of-the-art (SOTA) model by 1.35% and 2.61%, respectively. Besides, Spike-SLR only need 0.03mJ to process a sequence of event frames, achieving a 99.27% reduction in power consumption compared to the SOTA model. Code is available at https://github.com/Arktis2022/Spike-SLR.

1 Introduction

According to the latest data from World Federation of the Deaf, there are 70 million deaf people around the world using over 200 sign languages [23]. However, learning sign language is difficult and time-consuming, thus creating communication barriers to deaf people [11]. To address this issue, Sign Language Recognition (SLR) has been extensively researched. RGB video is the most commonly used input modality for SLR [16, 22, 27, 35, 36, 37]. However, the recognition results of RGB-based SLR methods are inevitably influenced by the motion blur inherent to RGB cameras and static background noise [38, 39].

*Xinxu Lin and Mingxuan Liu contributed equally to this work.

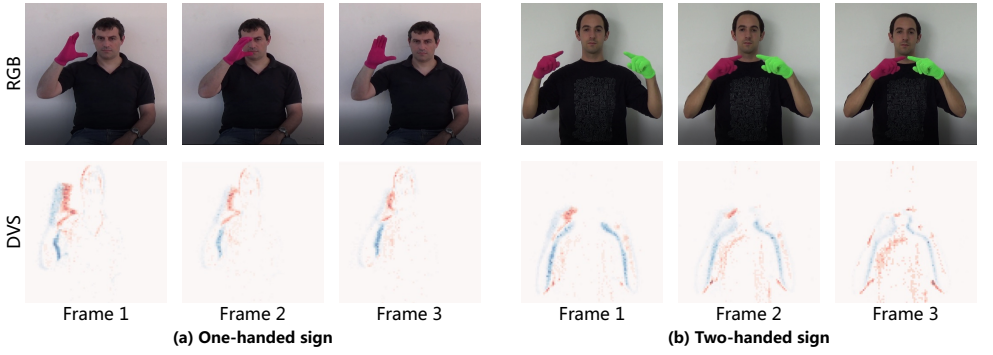

Figure 1: (a) Comparison of RGB video frames and DVS data frames for sign language Opaque(one-handed sign) (b) Comparison of RGB video frames and DVS data frames for sign language map(two-handed sign)

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 and DVS event frames is shown in Figure 1. The event camera features high temporal resolution, low latency, low power consumption, and a wide dynamic range [32], which can effectively address issues related to motion blur and static background noise. That is, event cameras hold significant advantages in the field of SLR. The current state-of-the-art (SOTA) approaches for event-based SLR involve first converting event streams into frame data, followed by processing using Artificial Neural Networks (ANNs) [3, 4, 5, 14, 26], 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 [20]. Compared to ANNs, the event-driven nature of SNNs significantly reduces energy consumption when running on neuromorphic chips [44, 45]. However, current SNN-based sign language recognition tasks still face challenges of lack of datasets and low recognition accuracy [28].

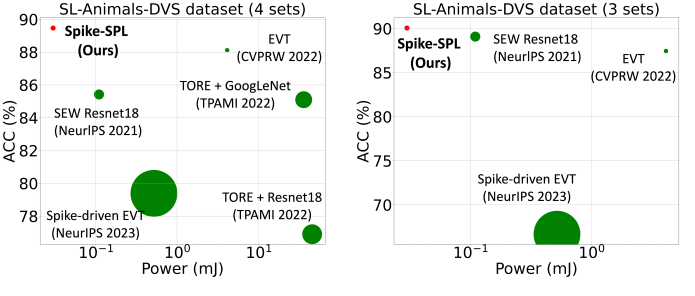

Figure 2: Accuracy vs inference energy of different neural methods implemented in Intel Stratix 10 TX [6] (for ANNs) or ROLLS [24] (for SNNs). The size of the markers denotes the number of parameters.

In this paper, in order to simultaneously reduce the power consumption and improve the accuracy in the event-based SLR, we propose a parallel spiking transformer model called Spike-SLR, which consists of a patch embedding (PE) block, parallel transformer blocks, a soft-attention block, and a classification head. As shown in Figure 2, we demonstrate that the proposed Spike-SLR outperforms other event-based SLR models by a significant margin, requiring less power consumption. And the main contributions of this paper are listed:

(1) We introduce the spiking transformer into SLR for the first time. And to improve the model's spatio-temporal attention to fine-grained hand features, we employ parallel spiking transformer, where multiple-layer perceptrons (MLPs) and simplified attention sub-modules (CB-S3A) are executed in parallel to improve efficiency.

(2) We firstly introduce the soft attention mechanisms into SNNs and employ a soft-attention block in our model to extract key regions from the input event streams.

(3) Experiments on the public datasets SL-Animals-DVS-4sets [33], SL-Animals-DVS-3sets [33], and the DVS datasets N-LSA64-Both and N-LSA64-Right converted using the v2e [13] method, demonstrate that Spike-SLR achieves better recognition accuracy compared to SOTA ANN method EVT [26]. And Spike-SLR only needs 0.03mJ to process an event frame sequence compared to the EVT which needs 4.13 mJ.

2 Related work

2.1 Event-based Sign Language Recognition

Event cameras are ideal for SLR due to their motion detection abilities, sparking increased research interest. However, event-based SLR still face the problem of lack of datasets, the only public event-based SLR dataset is SL-Animals-DVS [33], featuring 19 gestures. Although Qi Shi et al. created the synthetic N-WLASL [28] dataset, it remains private. SL-Animals-DVS is firstly tested on three kinds of SNNs named SLAYER [29], STBP [41], and DECOLLE [15], separately, where the test accuracy is all below 75%. To enhance the test accuracy, recent studies mainly use ANN methods. Laure Acin et al. introduced VK-SITS [2], a new event data representation, using a ResNet18 network, which outperformed other methods like TORE [9] and SITS [21]. Apart from that, Alberto Sabater et al. developed EVT [26], an efficient transformer model utilizing event data sparsity and becoming the SOTA method on the SL-Animals-DVS dataset.

2.2 Spiking Transformers

ANN-based transformers have achieved success in fields such as vision and natural language processing (NLP) [1, 10]. 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 [34] are incompatible with SNNs. Recently, research has increasingly focused on developing the spiking transformer, aimed at eliminating multiplication operations in SA to reduce computational complexity. Zhou et al. [46] 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. [43] 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 et al. [40] 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 [9, 17, 43].

3 Method

The proposed Spike-SLR applies the spiking transformer to sign language recognition tasks. We utilize the SNNs algorithm provided in the SpikingJelly platform [8], employing the Leak Integrate and Fire (LIF) [31] neural model for constructing the spiking neuron layers.

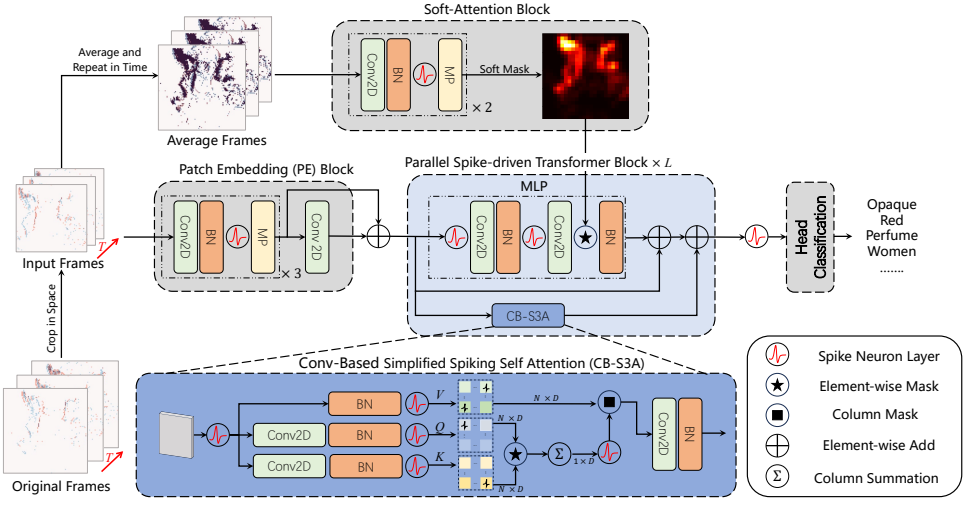

Figure 3: Framework of Spike-SLR. We follow the network structure in [43]. It consists of an SNN-based patch embedding (PE) block, several parallel spike-driven transformer blocks, a soft-attention block, and a SNN-based predictor head.

3.1 Overall architecture

Figure 3 shows the structure of Spike-SLR, which consists of four main components: patch embedding (PE) block, parallel spike-driven transformer block, soft-attention block and a classification head. The PE block is utilized to extract spatio-temporal representations from the input DVS frames, while the CB-S3A module in the transformer and the spike firing rate map in soft-attention mask block guide the model's attention towards key features. The final predictor head is responsible for mapping these features to possiable 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 layers, three SNN layers and two max pooling layers, downsampling 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 shape as S_{PE}. Therefore, the SNN-based PE block can be written as follows:

$$I = \mathrm{PR}(I_0) \quad I_0 \in \mathbb{R}^{T_0 \times 2 \times H_0 \times W_0}, I \in \mathbb{R}^{T \times 2 \times H \times W} \tag{1}$$

$$S_{PE} = \mathrm{PE}(I) \quad S_{PE} \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}} \tag{2}$$

$$S_0 = I_{PE} + S_{PE} \quad S_0 \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}} \tag{3}$$

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-SLR, 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 Soft-attention block 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:

$$S_l = \text{CB-S3A}(S_{l-1}) + \mathrm{MLP}(S_{l-1}) + S_{l-1} \quad S_l \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}}, l = 0...L \tag{4}$$

$$S_E = \mathcal{SN}(S_L) \quad S_E \in \mathbb{R}^{T \times D \times \frac{H}{4} \times \frac{W}{4}} \tag{5}$$

$$Y = SCH(S_E) \tag{6}$$

3.2 Soft-attention masks

DVS data can be influenced by various sources of noise, such as environmental background. As neural networks deepen, some noisy data may be amplified, causing the model to focus more on irrelevant features. And during the presentation of sign language, the frequencies and quantities of spike signals generated by different body parts are uneven. Inspired by [18], we insert attention blocks into our model to minimize the negative impact of background noise while allowing the model to focus more on the target area and local features. In order to take note of the difference among different body parts, soft attention mask is applied to assign higher weights to pixels with stronger spike signals, while it is also the bridge between Soft-attention appearance and the backbone network. Unlike the data processing operations

performed in the PE block, we first perform a sum-average-repeat (SAR) operation on the data in the soft 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 attention block model. 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 3.

3.3 Parallel spike-driven transformer

In the previous spiking Transformer architecture [42, 43, 46], 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:

$$U_{\text{out}} = \alpha_{\text{FF}}\hat{U} + \beta_{\text{FF}}\text{MLP}(\mathcal{SN}(\hat{U})) \tag{7}$$

$$\hat{U} = \alpha_{\text{SA}}U_{\text{in}} + \beta_{\text{SA}}\text{SA}(\mathcal{SN}(U_{\text{in}})) \tag{8}$$

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:

$$S_{\text{out}} = \alpha_{comb}S_{\text{in}} + \beta_{\text{FF}}\text{MLP}(\mathcal{SN}(S_{\text{in}})) + \beta_{\text{SA}}\text{SA}(\mathcal{SN}(S_{\text{in}})) \tag{9}$$

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:

$$MASK\left(Q_S, K_S, V_S\right) = g\left(Q_S, K_S\right) \otimes V_S = \mathcal{SN}\left(SUM_C\left(Q_S \otimes K_S\right)\right) \otimes V_S \tag{10}$$

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 Experiments

4.1 Dataset

To evaluate the Spike-SLR model, we use the public dataset SL-Animals-DVS [33] and the N-LSA64 dataset which is transformed from LSA64 [25] dataset using v2e [13] method. 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. As for the N-LSA64, It contains 3200 DVS videos in which 10 non-expert subjects performed 5 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 partition the right-hand-only dataset N-LSA64-Right.

4.2 Implementation details

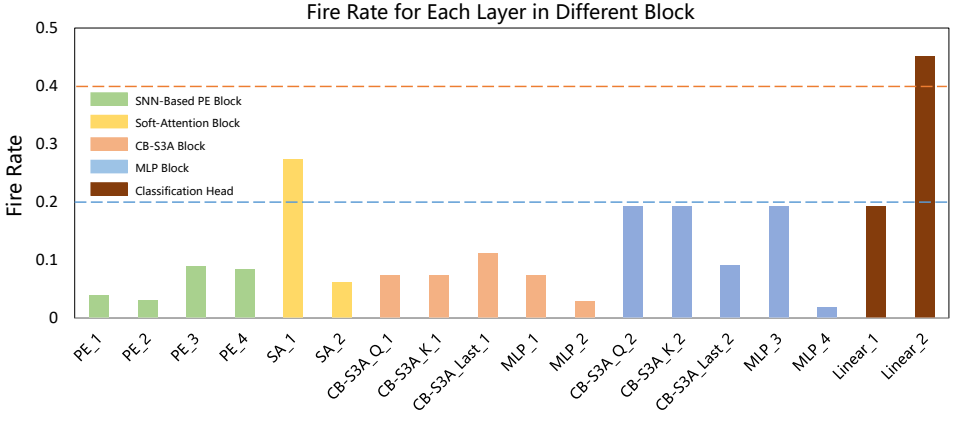

Figure 4: Fire rate of each block in Spike-SLR

We set the number of parallel spike-driven transformer block $L = 2$ in Spike-SLR. The whole framework is optimized with AdamW [19] optimizer, in a single NVIDIA GeForce RTX 3090. The sample length, time step, and learning rate is set as 500ms, 10 and $4\times e^{-3}$ respectively. We set the batch size to 32 and trained for 240 epochs using the 1cycle learning rate policy [30]. 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, 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

On the SL-Animals-DVS dataset, we compare our proposed model with existing ANN models [2, 4, 21, 26, 47], and SNN models [7, 15, 29, 41, 43]. Additionally, we replace the backbone network in EVT [26] with the Spike-Driven Transformer block [43] to obtain Spike-Evt and conduct model training for comparative analysis. Our experimental results on SL-Animals-DVS are given in Table 1, from which we can see that Spike-SLR is 1.35% and 2.61% higher than EVT [26] on the dataset SL-Animals-DVS-4sets and SL-Animals-DVS-3sets, respectively. And compared to the SNN method SEW Resnet18 [7], the Spike-SLR improves the accuracy of 4.05% and 0.97%, respectively.

On the N-LSA64-Both and N-LSA64-Right datasets, we employ the same sampling and training strategies to train ANN model EVT [26] and SNN models STBP [41] and Spike-

Table 1: Classification accuracy in the SL-Animals-DVS dataset. Red and **bold** indicate the best and second best performance.

Model	Method	Time Step	Sample Length	SL-Animals-DVS	
				4 sets	3 sets
TORE+ GoogLeNet [4]	ANN	\	\	85.10%	\
TORE+ ResNet18 [4]		\	\	76.90%	\
VoxelGrid+ ResNet18 [47]		\	\	**89.02%**	\
SITS + ResNet18 [21]		\	\	78.47%	\
VK-SITS+ ResNet18 [2]		\	\	79.26%	\
EVT [26]		\	504ms	88.12%	87.45%
SLAYER [29]	SNN	300	1500ms	54.30%	61.41%
STBP [41]		50	1500ms	64.97%	71.47%
DECOLLE [15]		500	500ms	62.19%	62.19%
SEW Resnet18 [7]		16	\	85.42%	**89.09%**
Spike-driven EVT [43]		11	504ms	79.39%	66.67%
Spike-SLR (Ours)	SNN	10	500ms	**89.47%**	**90.06%**

Table 2: Classification accuracy in the N-LSA64 dataset.

Model	Method	Time Step	Sample Length	N-LSA64	
				Both	Right
EVT [26]	ANN	\	504ms	**84.06%**	82.14%
STBP [41]	SNN	50	1500ms	59.69%	57.86%
Spike-driven EVT [43]		11	504ms	72.66%	**82.62%**
Spike-SLR (Ours)	SNN	10	500ms	**84.69%**	**86.90%**

EVT. The test results are shown in Table 2, from which we can find that Spike-SLR improves the accuracy of 0.63% compared to the EVT [26] on the N-LSA64-Both dataset while of 4.28% compared to the Spike-driven EVT [43] on the N-LSA64-Right dataset. In summary, Spike-SLR obtains the SOTA results in event-based SLR.

4.4 Energy consumption analysis

We use the same energy efficiency calculation scheme as in [12]. The energy consumption is 12.5 pJ for each floating-point operation (FLOP) and is 77 fJ for each synaptic operation (SOP). Figure 4 shows the average spike fire rate of convolutional layers in different blocks and linear layers in the classification head. We find that the spike fire rate (SFR) of the input tensor in different transformer blocks shows an upward trend but consistently stays below 0.2, which is because of the cumulative effect of input across layers. On the other hand, in the continuous convolutional layers of the MLP and Soft-Attention blocks, the spike rate shows a downward trend, indicating that some less significant features are filtered out as the network deepens. As shown in Table 3, the Spike-SLR processes DVS frame data with a

Table 3: Computational complexity comparisons of SLR methods.

Model	Method	#Params.	FLOPs/SOPs	Power/mJ
TORE + ResNet18 [4]	ANN	11.69 M	3.66 G	45.75
TORE + GoogLeNet [4]	ANN	8.46 M	2.88 G	36.00
EVT [26]	ANN	0.50 M	0.33 G	4.13
Spike-driven EVT [43]	SNN	66.34 M	6.77 G	0.52
SEW Resnet18 [7]	SNN	2.92 M	1.41 G	0.11
Spike-SLR (Ours)	SNN	0.70 M	0.44 G	**0.03**

Table 4: Spike-SLR accuracy with different time step.

Time Steps	1	5	10	15	20
ACC	77.63%	88.16%	89.47%	87.28%	89.04%

spatial size of 96x96 and a time step of 10 with only 0.03mJ of power consumption. This represents a 99.27% energy reduction compared to EVT and is substantially lower than that of other baseline models.

4.5 Ablation study

In this section, we analyze the impact of hyperparameters and the key components of Spike-SLR. Experiments are conducted on the SL-Animals-DVS-4sets dataset. With a fixed total sample length of 500ms, different time steps are set to investigate the impact of the number of input event frames and transformer blocks on the model results.

As shown in Table 4 and Table 5, with the number of time steps and the number of Transformer Blocks increasing, the test accuracy of the model does not change significantly, but too few time steps or blocks can lead to a significant decrease in accuracy. Specifically, setting the time step to 10 and the number of blocks to 2 achieve the highest accuracy of 89.47%. Reducing the time step to 1 decreased the accuracy to 77.63%, and setting the number of blocks to 1 result in an accuracy of 84.65%. In addition, The experiments validate the rationality of the parallel structure and the Soft-Attention appearance used in the Spike-SLR model. As shown in Table 6, using both parallel transformers and soft attention simultaneously yields the best accuracy, at 89.47%. Employing only the parallel transformer achieves an accuracy of 84.21%, while using solely soft attention results in an accuracy of 84.65%. Overall, the Spike-SLR model structure proposed by us achieves best results.

5 Conclusion

In this paper, we propose a spiking transformer, coined as Spike-SLR for event-based sign language recognition, to adaptively emphasis on local spatial features as well as temporal

Table 5: Spike-SLR accuracy with different number of parallel transformer block.

Blocks	1	2	3	4	5
ACC	84.65%	89.47%	88.60%	88.60%	89.04%

Table 6: Spike-SLR accuracy for different architecture.

	Parallel Transformer block	Serial Transformer block
w/o soft attention	84.21%	86.40%
with soft attention	89.47%	84.65%

features. Spike-SLR outperforms existing methods upon SL-Animals-DVS and N-LSA64 datasets in accuracy and energy consumption. Specifically, Spike-SLR improves the accuracy of 1.35% and 2.61% respectively compared to the SOTA ANN model while reducing power consumption of 99.27% on the two SL-Animals-DVS datasets. It demonstrates the applicability of spiking transfomer for SLR and can be further applied to human-machine interaction and edge sign language recognition devices.

Acknowledgments

This work was supported by the NationalScience and Technology Major Project from the Minister of Science and Technology, China, under Grant 2018AAA0103100; and in part by the National Natural Science Foundation of China under Grant 92164110 and 62334014. (Corresponding author: Hong Chen.)

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Laure Acin, Pierre Jacob, Camille Simon-Chane, and Aymeric Histace. Vk-sits: a robust time-surface for fast event-based recognition. In *2023 Twelfth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2023.

[3] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7388–7397, 2017. doi: 10.1109/CVPR.2017.781.

[4] R Wes Baldwin, Ruixu Liu, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Time-ordered recent event (tore) volumes for event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2519–2532, 2022.

[5] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. 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*, page 136–152, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58564-8. doi: 10.1007/978-3-030-58565-5_9. URL https://doi.org/10.1007/978-3-030-58565-5_9.

[6] Intel Corporation. Intel stratix 10 tx device overview. https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/hb/stratix-10/s10_tx_overview.pdf, 2023. Accessed: 2023-12-10.

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

[8] Wei Fang, Yanqi Chen, Jianhao Ding, Zhaofei Yu, Timothée Masquelier, Ding Chen, Liwei Huang, Huihui Zhou, Guoqi Li, and Yonghong Tian. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40):eadi1480, 2023.

[9] Lingyue Guo, Zeyu Gao, Jinye Qu, Suiwu Zheng, Runhao Jiang, Yanfeng Lu, and Hong Qiao. Transformer-based spiking neural networks for multimodal audio-visual classification. *IEEE Transactions on Cognitive and Developmental Systems*, 2023. URL https://api.semanticscholar.org/CorpusID:264480751.

[10] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[11] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2529–2539, June 2023.

[12] Yangfan Hu, Huajin Tang, and Gang Pan. Spiking deep residual networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):5200–5205, 2023. doi: 10.1109/TNNLS.2021.3119238.

[13] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1312–1321, 2021. doi: 10.1109/CVPRW53098.2021.00144.

[14] Simone Undri Innocenti, Federico Becattini, Federico Pernici, and Alberto Del Bimbo. Temporal binary representation for event-based action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10426–10432, 2021. doi: 10.1109/ICPR48806.2021.9412991.

[15] Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:515306, 2020.

[16] Ahmet Kındıroglu, Oğulcan Özdemir, and Lale Akarun. Aligning accumulative representations for sign language recognition. *Machine Vision and Applications*, 34, 12 2022. doi: 10.1007/s00138-022-01367-x.

[17] Mingxuan Liu, Jiankai Tang, Haoxiang Li, Jiahao Qi, Siwei Li, Kegang Wang, Yuntao Wang, and Hong Chen. Spiking-physformer: Camera-based remote photoplethysmography with parallel spike-driven transformer. *ArXiv*, abs/2402.04798, 2024. URL https://api.semanticscholar.org/CorpusID:267522897.

[18] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19400–19411. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e1228be46de6a0234ac22ded31417bc7-Paper.pdf.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

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

[21] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, and Vincent Lepetit. 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*, pages 10245–10254, 2019.

[22] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Using motion history images with 3d convolutional networks in isolated sign language recognition. *IEEE Access*, 10: 18608–18618, 2022. doi: 10.1109/ACCESS.2022.3151362.

[23] J. Murray. World federation of the deaf, 2018. http://wfdeaf.org/our-work/, Last accessed on 2024-05-08.

[24] Ning Qiao, Hesham Mostafa, Federico Corradi, Marc Osswald, Fabio Stefanini, Dora Sumislawska, and Giacomo Indiveri. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses. *Frontiers in neuroscience*, 9:123487, 2015.

[25] Franco Ronchetti, Facundo Manuel Quiroga, César Estrebou, Laura Lanzarini, and Alejandro Rosete. Lsa64: An argentinian sign language dataset. *ArXiv*, abs/2310.17429, 2023. URL https://api.semanticscholar.org/CorpusID:64745485.

[26] Alberto Sabater, Luis Montesano, and Ana Cristina Murillo. Event transformer. a sparse-aware solution for efficient event data processing. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2676–2685, 2022. URL https://api.semanticscholar.org/CorpusID:248006332.

[27] Xiaolong Shen, Zhedong Zheng, and Yi Yang. Stepnet: Spatial-temporal part-aware network for isolated sign language recognition. *ACM Trans. Multimedia Comput. Commun. Appl.*, apr 2024. ISSN 1551-6857. doi: 10.1145/3656046. URL https://doi.org/10.1145/3656046.

[28] Qi Shi, Zhongfu Ye, Jin Wang, and Yueyi Zhang. Qisampling: An effective sampling strategy for event-based sign language recognition. *IEEE Signal Processing Letters*, 2023.

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

[30] Leslie N. Smith and Nicholay Topin. Super-convergence: very fast training of neural networks using large learning rates. In *Defense + Commercial Sensing*, 2018. URL https://api.semanticscholar.org/CorpusID:260552651.

[31] RB Stein and Alan Lloyd Hodgkin. The frequency of nerve action potentials generated by applied currents. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 167(1006):64–86, 1967.

[32] L Su, F Yang, XY Wang, CD Guo, LL Tong, and Q Hu. A survey of robot perception and control based on event camera. *Acta Autom. Sin*, 48:1869–1889, 2022.

[33] Ajay Vasudevan, Pablo Negri, Camila Di Ielsi, Bernabe Linares-Barranco, and Teresa Serrano-Gotarredona. Sl-animals-dvs: event-driven sign language animals dataset. *Pattern Analysis and Applications*, pages 1–16, 2022.

[34] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:13756489.

[35] Manuel Vázquez-Enríquez, José L. Alba-Castro, Laura Docío-Fernández, and Eduardo Rodríguez-Banga. 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)*, pages 3457–3466, 2021. doi: 10.1109/CVPRW53098.2021.00385.

[36] Fei Wang, Libo Zhang, Hao Yan, and Shuai Han. Tim-slr: a lightweight network for video isolated sign language recognition. *Neural Comput. Appl.*, 35(30):22265–22280, aug 2023. ISSN 0941-0643. doi: 10.1007/s00521-023-08873-7. URL https://doi.org/10.1007/s00521-023-08873-7.

[37] Pichao Wang, Wanqing Li, Zhimin Gao, Yuyao Zhang, Chang Tang, and Philip Ogunbona. 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)*, pages 416–425, 2017. doi: 10.1109/CVPR.2017.52.

[38] Qinyi Wang, Yexin Zhang, Junsong Yuan, and Yilong Lu. Space-time event clouds for gesture recognition: From rgb cameras to event cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1826–1835, 2019. doi: 10.1109/WACV.2019.00199.

[39] Yanxiang Wang, Xian Zhang, Yiran Shen, Bowen Du, Guangrong Zhao, Lizhen Cui, and Hongkai Wen. Event-stream representation for human gaits identification using deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3436–3449, 2022. doi: 10.1109/TPAMI.2021.3054886.

[40] Ziqing Wang, Yuetong Fang, Jiahang Cao, Qiang Zhang, Zhongrui Wang, and Renjing Xu. Masked spiking transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1761–1771, 2023.

[41] Yujie Wu, Lei Deng, Guoqi Li, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12: 323875, 2018.

[42] Man Yao, Jiakui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. *arXiv preprint arXiv:2404.03663*, 2024.

[43] Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] Jilin Zhang, Mingxuan Liang, Jinsong Wei, Shaojun Wei, and Hong Chen. A 28nm configurable asynchronous snn accelerator with energy-efficient learning. In *2021 27th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC)*, pages 34–39. IEEE, 2021.

[45] Jilin Zhang, Dexuan Huo, Jian Zhang, Chunqi Qian, Qi Liu, Liyang Pan, Zhihua Wang, Ning Qiao, Kea-Tiong Tang, and Hong Chen. 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. In *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 21–23. IEEE, 2023.

[46] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Liuliang Yuan. Spikformer: When spiking neural network meets transformer. *ArXiv*, abs/2209.15425, 2022. URL https://api.semanticscholar.org/CorpusID:252668422.

[47] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019.

[48] Shihao Zou, Yuxuan Mu, Xinxin Zuo, Sen Wang, and Chao Li. Event-based human pose tracking by spiking spatiotemporal transformer. *ArXiv*, abs/2303.09681, 2023. URL https://api.semanticscholar.org/CorpusID:257622694.