STAA-SNN: Spatial-Temporal Attention Aggregator for Spiking Neural Networks

Tianqing Zhang[1*] Kairong Yu[1*] Xian Zhong[3] Hongwei Wang[1†] Qi Xu[2†] Qiang Zhang[2]
[1] Zhejiang University [2] Dalian University of Technology [3] Wuhan University of Technology
[1]{zhangtianqing,22271086}@zju.edu.cn [1]hongweiwang@intl.zju.edu.cn
[2]{xuqi,zhangq}@dlut.edu.cn [3]zhongx@whut.edu.cn

Abstract

Spiking Neural Networks (SNNs) have gained significant attention due to their biological plausibility and energy efficiency, making them promising alternatives to Artificial Neural Networks (ANNs). However, the performance gap between SNNs and ANNs remains a substantial challenge hindering the widespread adoption of SNNs. In this paper, we propose a Spatial-Temporal Attention Aggregator SNN (STAA-SNN) framework, which dynamically focuses on and captures both spatial and temporal dependencies. First, we introduce a spike-driven self-attention mechanism specifically designed for SNNs. Additionally, we pioneeringly incorporate position encoding to integrate latent temporal relationships into the incoming features. For spatial-temporal information aggregation, we employ step attention to selectively amplify relevant features to variant steps. Finally, we implement a time-step random dropout strategy to avoid local optima. The framework demonstrates exceptional performance across diverse datasets and exhibits strong generalization capabilities. Notably, STAA-SNN achieves state-of-the-art results on neuromorphic datasets CIFAR10-DVS of 82.10% and with performances of 97.14%, 82.05% and 70.40% on the static datasets CIFAR-10, CIFAR-100 and ImageNet, respectively. Furthermore, this model exhibits improved performance ranging from 0.33% to 2.80% with fewer time steps.

1. Introduction

In recent years, Spiking Neural Networks (SNNs) have attracted significant attention due to their low energy consumption and biological interpretability[12, 17, 51, 57–59]. Inspired by the behavior of biological neurons, SNNs represent information using discrete binary spikes over multiple timesteps, making them well-suited for implementation on low-power neuromorphic hardware, offering a distinct advantage over traditional Artificial Neural Networks (ANNs)[2, 13, 19, 48, 63]. Among various spiking neuron models, the Leaky Integrate-and-Fire (LIF) model[1] is widely used in SNNs for its linear differentiable [37, 38]. Although previous works[17, 34, 73] have developed deep SNNs to accelerate model convergence, the multi-timestep computations inherent in SNNs significantly increase both training and inference times, leading to high latency. This limitation hinders the potential for efficient algorithm design. Consequently, a performance gap still exists between directly trained large-scale SNNs and ANNs in pattern recognition tasks, underscoring the need for novel approaches that integrate insights from neuroscience with conventional deep learning to improve both the efficiency and accuracy of SNNs. Inspired by the human visual and cognitive systems, attention mechanisms regulate neural activity and connectivity, enhancing responses to focal stimuli. This enhancement is reflected in the increased firing rates and improved synchronization of neurons in relevant brain regions. Additionally, this mechanism influences the strength and efficiency of neural connections, thereby facilitating the flow of information within and between different brain areas. In the field of deep learning, attention mechanisms enable neural networks to selectively focus on relatively important input information, leading to significant advancements in various domains such as natural language processing (NLP) and computer vision (CV). Compared to traditional architectures [5, 6, 44], these mechanisms have consistently demonstrated superior performance.

In this paper, we explore the integration of attention mechanisms into deep SNNs. Although SNNs are biologically inspired models, their current implementations exhibit significant limitations due to inadequate biological fidelity. For instance, the heterogeneity among spiking neurons and the ability to independently transmit spatial and temporal information are overlooked, with membrane parameters being uniformly and permanently set across dimensions and layers in neural networks. Neuroscience research [3, 53] has

*Equal contribution.
†Corresponding authors.

highlighted the critical role of synaptic homeostatic plasticity mechanisms in neuronal activity, as well as the regulatory function of glial cells in information transmission between neurons. Enhancing the alignment of SNN models with these biological processes has the potential to improve their overall performance. Therefore, we propose allowing different network layers to exhibit varying degrees of information transmission permeability.

We introduce a self-attention mechanism to capture spatial correlations and the significance of information, while integrating positional encoding to enhance spatial information and potentially learn temporal features. In traditional SNNs, input features are typically aggregated through simple addition, which can introduce significant noise into the synthesized features. To address this, we propose a Step Attention mechanism that aggregates and reconstructs features at each time step, thereby improving the quality of information propagation. Additionally, we introduce a novel Time-Step Random Dropout (TSRD) strategy to prevent the network from falling into local optima, thereby enhancing its robustness and accelerating training. Our main contributions are summarized as follows:

- Inspired by neuroscience and attention mechanisms, we propose a novel Spatio-Temporal Attention Aggregator for Spiking Neural Networks (STAA-SNN) that integrates self-attention, positional encoding, and step attention into a unified framework, enabling SNNs to efficiently extract features across temporal and spatial domains while maintaining stable training.
- We introduce a Time-Step Random Dropout (TSRD) strategy to accelerate training and enhance model generalization by preventing premature convergence and enabling optimal performance.
- We evaluate the performance, convergence, and sparse spiking activity of STAA-SNN across multiple benchmark datasets, demonstrating that our method achieves state-of-the-art or superior results with fewer timesteps, highlighting its efficiency and effectiveness.

2. Related Works

2.1. Training of Deep SNNs

Training deep SNNs generally follows two primary approaches: ANN-to-SNN conversion and direct SNN training. The ANN-to-SNN conversion method aims to transfer knowledge from traditional ANNs to SNNs, but it often encounters challenges such as minimizing accuracy degradation while optimizing inference latency and energy efficiency[27, 28, 55]. Moreover, this approach struggles to fully leverage the spatiotemporal dynamics inherent in SNNs due to the lack of temporal encoding in ANNs. Direct SNN training is categorized into unsupervised and supervised learning. Unsupervised methods rely on biologically plausible learning rules such as Hebbian learning[30] and Spike-Timing-Dependent Plasticity (STDP)[7]. However, these approaches are generally not suitable for deep SNNs or large-scale datasets. On the other hand, supervised learning integrates backpropagation with surrogate gradients[17, 38, 46, 58], making applicability increasingly viable[60, 61]. Several methods have been proposed to enhance direct SNN training. For instance, MPBN[26] incorporates a batch normalization layer after membrane potential updates to stabilize training, while MPD-AL[72] introduces an alternative learning paradigm that adapts synaptic modifications based on membrane potential traces. MS-ResNet[34] reorganizes Vanilla ResNet layers to better align with SNNs, improving feature extraction and gradient propagation. Furthermore, IM-Loss[22] optimizes information maximization loss to increase spike information entropy. Recently, RSNN[62] introduced a Convolutional LSTM-based approach that preserves temporal correlations, enhancing spike-based data processing in SNNs.

2.2. Attention Mechanism in SNNs

Attention mechanisms dynamically assign varying levels of importance to different input components, thereby facilitating efficient processing and feature extraction. Since their introduction in 2014, attention mechanisms have become integral to enhancing deep learning models and have been widely applied to traditional ANNs[32]. In deep learning, attention mechanisms are primarily used in two ways: as a basic paradigm for conducting meta-operator such as self-attention [54] or as an auxiliary enhancement modules [33]. Also these two approaches can be combined. Recently, integrating attention mechanisms into SNNs has gained traction as a means to enhance their performance. Unlike traditional ANNs, where attention typically refines spatial feature representations, SNN-based attention mechanisms must consider both spatial and temporal information. Early work by [47] introduced a computational model that bridges human selective attention and conscious awareness. Kunduη [36] utilize attention mechanisms to compress SNN models. Temporal-wise attention was first introduced in TA-SNN[64], laying the foundation for subsequent developments in SNN-specific attention mechanisms. MA-SNN [66] proposed temporal-channel-spatial attention for SNNs, surpassing models focused on single-dimensional information. TCJA-SNN[76] introduced a temporal-channel joint attention mechanism, while SCTFA [8] applied temporal-channel-spatial attention to SNNs with promising results. Moreover, STSC-SNN [71] incorporated temporal convolution and attention mechanisms to implement synaptic filtering and gating functions, demonstrating the versatility of attention mechanisms in enhancing SNN capabilities. IM-LIF [43] adopted temporal-wise attention to improve the updating of GRUs and LSTMs

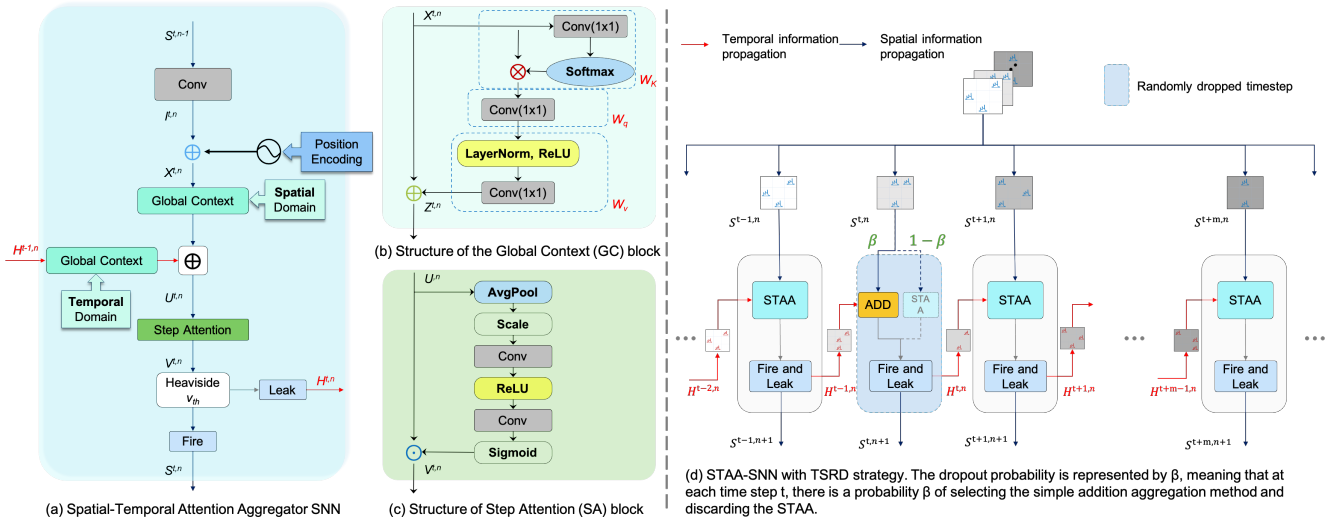

Figure 1. Overview of the STAA-SNN Architecture and the TSRD strategy.

within SNNs. More recently, researchers have explored incorporating SNNs into Transformer architectures, achieving promising results on multiple benchmark datasets[41, 52]. These researches reveal that attention mechanisms can significantly improve the performance of SNN. However, there remains a need for more specialized and high-performance attention modules tailored explicitly for SNN architectures. Addressing this gap can further enhance the efficiency and robustness of SNNs in real-world applications.

3. Methodology

In this section, we will first provide a brief overview of the Leaky Integrate-and-Fire (LIF) model and the training process. Following that, we introduce the proposed Spatial-Temporal Attention Aggregator. Lastly, we outline the timestep random dropout strategy for SNN.

3.1. Leaky Integrate-and-Fire Model

The spiking neuron is a fundamental component of SNNs. In this work, we adopt the Leaky Integrate-and-Fire (LIF) neuron due to its efficiency and simplicity. The discrete-time, iterative update rule of the LIF model is given by:

$$V^{t,n} = H^{t-1,n} + \frac{1}{\tau}[I^{t-1,n} - (H^{t-1,n} - V_{\text{reset}})], \quad (1)$$

$$S^{t,n} = \Theta(V^{t,n} - v_{\text{th}}), \quad (2)$$

$$H^{t,n} = V_{\text{reset}} \cdot S^{t,n} + V^{t,n} \odot (1 - S^{t,n}). \quad (3)$$

Here, $H^{t-1,n}$ denotes the membrane potential after a spike trigger at the previous timestep, while $I^{t,n}$ and $V^{t,n}$ represent the input and membrane potential of the n-th layer at timestep t, respectively. The threshold v_{th} determines whether a spike is emitted, producing the binary spike sensor $S^{t,n}$. The Heaviside step function Θ outputs 1 if $x \geq 0$ and 0 otherwise.

Given the similarity between iterative LIF dynamics and Recurrent Neural Networks (RNNs), Spatio-Temporal Backpropagation (STBP) [57, 58] enables direct SNN training via Backpropagation Through Time (BPTT):

$$\frac{\partial L}{\partial W^n} = \sum_t \left(\frac{\partial L}{\partial S^{t,n}} \cdot \frac{\partial S^{t,n}}{\partial V^{t,n}} + \frac{\partial L}{\partial V^{t+1,n}} \cdot \frac{\partial V^{t+1,n}}{\partial V^{t,n}} \right) \frac{\partial V^{t,n}}{\partial W^n}. \quad (4)$$

However, the term $\frac{\partial S^{t,n}}{\partial V^{t,n}}$, which represents the derivative of the spike function, is non-differentiable. To address this, prior works introduce Surrogate Gradients(SG), commonly approximated using a rectangular function:

$$\frac{\partial S^t}{\partial V^t} = \frac{1}{a} \cdot \text{sign}(|V^t - V_{\text{th}}| < \frac{a}{2}), \quad (5)$$

where a is a hypeparameter and set to 1. This approximation provides a gradient of 1 within the range $V_{\text{th}} - 0.5 \leq V^t \leq V_{\text{th}} + 0.5$, and otherwise 0.

Neuromorphic-Inspired Adaptive LIF Design. Examining Eq. (1), it is evident that keeping V_{reset} and τ constant results in fixed weight coefficients for H and I, which remain unchanged across timestep t and layer n. This simplification deviates from biological neural processing, where neuronal excitability dynamically adapts to stimuli. To bridge this gap, we propose an adaptive modification to the membrane potential update rule:

$$V^{t,n} = M \odot H^{t-1,n} + N \odot I^{t-1,n}. \quad (6)$$

Here, M and N are coefficient matrices that dynamically adjust through iterative learning, regulating activation and

inhibition probabilities across layers. This biologically inspired adaptation enhances information processing efficiency and robustness in SNNs.

Global Context (GC) Block. The Global Context (GC) Block serves as a lightweight alternative to the self-attention mechanism used in Transformers, enabling the integration of self-attention into conventional convolution-based deep learning architectures without the computational overhead of full Transformer models. This makes it possible to enhance feature learning in non-Transformer structures while maintaining efficiency and precision. In this paper, we simulate the self-attention operation using a GC block [9], which consists of three 1×1 convolutional kernels corresponding to key, query, and value. As depicted in Fig. 1(b), the spatial spiking input tensor of the n-th layer at the t-th timestep is defined as $X^{t,n} \in \mathbb{R}^{C^n \times H^n \times W^n}$, where W^n and H^n denote the width and height of the feature map, and C^n represents the number of channels. Passing $X^{t,n}$ through the GC block yields three weight matrices for self-attention, namely W_k, W_q and W_v, defined as:

$$W_k = X^{t,n} \cdot \text{Sigmoid}(\text{Conv}_k(X^{t,n})), \tag{7}$$

$$W_q = \text{Conv}_q(W_k), \tag{8}$$

$$W_v = \text{Conv}_v(\text{ReLU}(\text{LN}(W_q))). \tag{9}$$

Before computing W_k, $X^{t,n}$ is reshaped from $C \times H \times W$ to $C \times HW$. The resulting matrices W_k, W_q, $W_v \in \mathbb{R}^{C \times 1 \times 1}$ encode self-attention weights, which are then used to refine the input feature map. Subsequently, the enhanced output is computed as:

$$Z^{t,n} = X^{t,n} \oplus W_v, \tag{10}$$

where $Z^{t,n} \in \mathbb{R}^{C,H,W}$ maintains the same dimensions as $X^{t,n}$. The refined feature map $Z^{t,n}$, in comparison to the $X^{t,n}$, improves the network's ability to focus on self-relevant and task-critical information while suppressing irrelevant signals. This approach facilitates the seamless transmission of enriched data flow, enabling the retention of valuable features for subsequent stages.

Position Encoding (PE) Block. Position encoding is widely used in NLP and sequence modeling to incorporate positional information into input data. While SNNs share structural similarities with RNNs and inherently process temporal information, their reliance on sparse binary spikes for information storage can lead to significant feature loss. This study introduces position encoding to enhance spatial features while enriching temporal relationships, thereby improving feature retention and strengthening temporal representations. The positional encoding is computed as follows:

$$X^{t,n} = I^{t,n} \oplus \text{Pos}_{t,n}, \tag{11}$$

where $\text{Pos}_{t,n}$ represents the position encoding at the t-th time step in the n-th layer, with $\text{Pos}_{t,n} \in \mathbb{R}^{T,C^n}$.

Step Attention (SA) Block. While previous modules primarily enhance feature processing from a spatial perspective, SNNs uniquely process information across both spatial and temporal. Unlike other types of network, SNNs offer an additional understanding of information sensitivity across different timesteps. Therefore, in this paper, we introduce Step Attention to tackle the temporal challenges inherent in SNNs. As illustrated in Fig. 1(c), let $U^{t,n} \in \mathbb{R}^{C^n \times H^n \times W^n}$ denote the spatially aggregated feature input at the n-th layer and the t-th timestep. The specific Step Attention mechanism is defined as:

$$V^{t,n} = U^{t,n} \odot \text{Sigmoid}\Big(\text{Conv}_2\Big(\text{ReLU}\Big(\text{Conv}_1 \big(\alpha \cdot \text{AvgPool}(U^{t,n})\big)\Big)\Big)\Big), \tag{12}$$

where α is a scaling factor that regulates the initial training speed for smoother convergence. In practice, α is typically set to a default value of 2.

LIF with Spatio-Temporal Attention Aggregator (STAA). The STAA seamlessly integrates with the LIF model, offering compatibility across various network structures without constraints. It adeptly harnesses the potential of spiking neurons, effectively optimizing the storage of discrete spike information while minimizing the loss of crucial data during transmission. As shown in Fig. 1(a), the neuronal update equation for the LIF model with STAA is formulated as:

$$U^{t,n} = GC_1(X^{t,n}) \oplus GC_2(H^{t-1,n}), \tag{13}$$

$$V^{t,n} = SA(U^{t,n}). \tag{14}$$

Where GC_1 and GC_2 represent Global Context Blocks for the input $X^{t,n}$ and the previous membrane state $H^{t-1,n}$, respectively. The remaining update equations follow the original LIF model.

3.2. Time Step Random Dropout (TSRD) Strategy

To further enhance training efficiency, we propose TSRD, an adaptive training strategy designed to prevent premature feature stabilization in deep SNNs. In deeper timesteps, spiking features may solidify too early during iterations, hindering the network from achieving optimal convergence. As illustrated in Fig. 1(d), TSRD randomly dropout the augmentation module at timestep $t \in (0, T)$ with a dropout probability β, and instead choose the most basic element-wise addition aggregation method. The TSRD technique accelerates training and helps the network escape local optima, thereby improving generalization performance.

	Method	Architecture	Timestep	Accuracy (%)	
				CIFAR-10	CIFAR-100
CNN-based	GLIF [70]	ResNet-19	2	94.44%	75.48%
		ResNet-19	4	94.85%	77.05%
	TET [16]	ResNet-19	2	94.16%	72.87%
		ResNet-19	4	94.44%	74.47%
	LSG [42]	ResNet-19	2	94.41%	76.32%
		ResNet-19	4	95.17%	76.85%
	PFA [14]	ResNet-19	2	95.6%	76.7%
		ResNet-19	4	95.71%	78.1%
	Diet-SNN [50]	ResNet-20	5	91.78%	64.07%
		VGG-16	5	93.85%	69.67%
	IM-loss [22]	VGG-16	5	93.85%	70.18%
	IM-LIF [43]	ResNet-19	3	95.29%	77.21%
	MPBN [26]	ResNet-19	1	96.06%	78.71%
		ResNet-19	2	96.47%	79.51%
		ResNet-19	4	96.52%	80.1%
		ResNet-20	1	92.22%	68.41%
		ResNet-20	2	93.54%	70.79%
		ResNet-20	4	94.28%	72.3%
Transformer-based	Spikformer[75]	Spikformer-4-25	4	93.94%	75.96%
	Spikingformer[74]	Spikingformer-4-25	4	94.77%	77.43%
	Spike-driven Transformer [69]	Transformer-2-512	4	95.60%	78.40%
CNN-based	**Ours**	VGG-13	2	94.70% $\pm$ 0.12%	75.16% $\pm$ 0.08%
		VGG-13	4	95.26% $\pm$ 0.09%	76.35% $\pm$ 0.11%
		ResNet-19	1	96.75% $\pm$ 0.09%	79.37% $\pm$ 0.10%
		ResNet-19	2	96.85% $\pm$ 0.11%	80.57% $\pm$ 0.11%
		ResNet-19	4	**97.14%** $\pm$ 0.10%	**82.05%** $\pm$ 0.10%
		ResNet-20	1	93.08% $\pm$ 0.09%	70.14% $\pm$ 0.07%
		ResNet-20	2	94.35% $\pm$ 0.08%	73.20% $\pm$ 0.12%
		ResNet-20	4	95.03% $\pm$ 0.11%	75.10% $\pm$ 0.12%

Table 1. Comparison results with SOTA methods on CIFAR-10/100.

4. Experiments

We first outlined the experimental setup, including dataset details and implementation specifics. We then compare our results with previous SOTA methods across multiple benchmarks. Next, we perform ablation studies to assess the effectiveness of STAA. Finally, we visualize the model's performance. Further details on datasets, hyperparameters, additional experiments, and computational efficiency analysis are provided in the **Supplementary Material**.

4.1. Experimental Setup

Datasets. We evaluate our method on five benchmark datasets. CIFAR-10/100[35] are standard image classification datasets with 50,000 training and 10,000 test images across 10 and 100 classes, respectively. ImageNet[15] is a large-scale dataset containing 1.2 million training, 50,000 validation, and 100,000 test images spanning 1,000 categories. For neuromorphic vision tasks, CIFAR10-DVS[39, 40, 45] consists of 10,000 event streams across 10 classes, with a 9:1 training-to-testing split. DVS128 Gesture[4] is an event-based gesture recognition dataset with 1,176 training and 288 test samples, featuring 11 gesture classes recorded from 29 subjects under three lighting conditions.

Implementation Details. In our experiments, we set the firing threshold v_{th} to 1, and initialize α and β to 2 and 0.1, respectively. In the PE block, we adopt a learnable position encoding method, initializing all values to zero. All experiments are implemented in PyTorch and conducted on an NVIDIA RTX 3090 GPU, except for ImageNet, which is

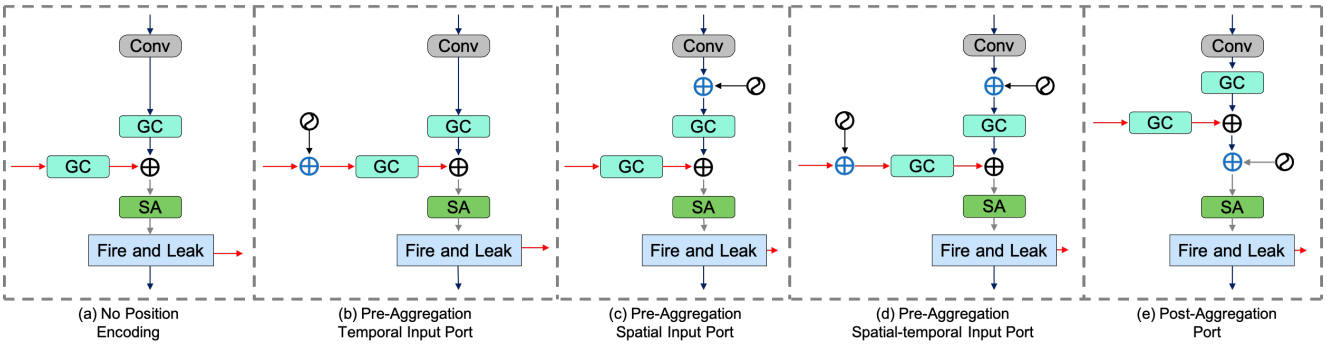

Figure 2. Position encoding locations in SNNs.

Method	Architecture	Timestep	Accuracy
STBP-tdBN [73]	ResNet34	6	63.72%
TET [16]	ResNet34	6	64.79%
RecDis-SNN [23]	ResNet34	6	67.33%
GLIF [70]	ResNet34	4	67.52%
IM-Loss [22]	ResNet18	6	67.43%
Real Spike [24]	ResNet18	4	63.68%
	ResNet34	4	67.69%
RMP-Loss [25]	ResNet18	4	63.03%
	ResNet34	4	65.17%
MPBN [26]	ResNet18	4	63.14%
	ResNet34	4	64.71%
SEW ResNet [17]	ResNet18	4	63.18%
	ResNet34	4	67.04%
Ours	ResNet18	4	**68.27%** ± 0.19%
	ResNet34	4	**70.40%** ± 0.15%

Table 2. Comparison of training based SNN SOTA on ImageNet.

trained on 8 NVIDIA RTX 4090 GPUs. The total training epochs are 500 for CIFAR-10, CIFAR-100, and CIFAR10-DVS, while ImageNet runs for 300 epochs.

4.2. Performance Comparison

Static Image Classification. We evaluate our model on three static datasets: CIFAR-10, CIFAR-100, and ImageNet. For CIFAR-10/100, STAA-SNNs are applied to both ResNet-like and VGG-like networks and tested at 1, 2, 3, and 4 timesteps. On ImageNet, the model is evaluated using ResNet-18 and ResNet-34 with 4 time steps. Our method achieves top accuracies of 97.14% on CIFAR-10, 82.05% on CIFAR-100, and 70.40% on ImageNet, surpassing previous state-of-the-art methods across various architectures. These improvements are attributed to the strong spatial-temporal attention mechanism and feature aggregation capabilities of STAA-SNNs, which fully leverage the potential performance of SNNs. Furthermore, our model achieves equivalent accuracy while requiring at least one fewer timestep compared to prior methods, improving efficiency without sacrificing performance.

Methods	Architecture	Timestep	Accuracy
CIFAR10-DVS			
IM-loss [22]	ResNet-19	10	72.60%
LSG [42]	ResNet-19	10	77.90%
MPBN [26]	ResNet-19	10	74.40%
MPBN [26]	ResNet-20	10	78.70%
TET [16]	VGGSNN	10	77.30%
IM-LIF [43]	VGG-13	10	80.50%
GLIF [70]	7B-wideNet	16	78.10%
STSA [56]	STS-Transformer	16	79.93%
SEW [17]	SEW-ResNet	16	74.4%
Spikeformer [41]	Spikeformer	16	80.9%
Ours	ResNet-20	16	81.90% ± 0.20%
	VGG-13	16	**82.10%** ± 0.20%
DVS128 Gesture			
STBP-tdBN [73]	ResNet-17	40	96.87%
SEW [17]	7B-Net	16	97.52%
PLIF [18]	PLIFNet	20	97.57%
MA-SNN [66]	5 layers SCNN	20	98.23%
ASA-SNN [67]	5 layers SCNN	20	97.70%
IM-LIF [43]	ResNet-19	40	97.33%
LIAF+TA [65]	TA-SNN-Net	60	98.61%
Spike-driven-Transformer [69]	Transformer-2-512	16	**99.30%**
Ours	VGG-13	16	98.61% ± 0.20%

Table 3. Comparison results with SOTA methods on CIFAR10-DVS and DVS128 Gesture.

Event-based Action Recognition. To further assess the spatial-temporal processing capabilities of STAA-SNNs,

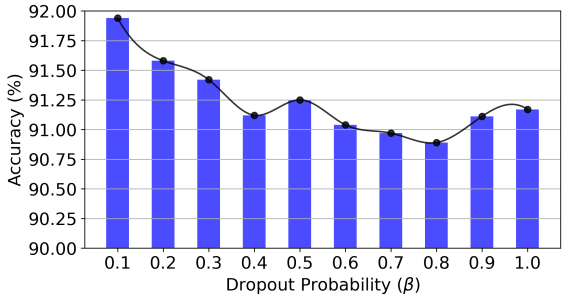

Figure 3. Distribution of accuracy with different dropout probability β in TSRD on CIFAR-10.

STAA Components			TSRD	Accuracy	Accuracy Gain
GC	PE	SA			
✗	✗	✗	✗	72.30%	-
✓	✗	✗	✗	73.22%	+0.92%
✓	✓	✗	✗	73.79%	+1.49%
✓	✓	✗	✓	73.96%	+1.66%
✓	✓	✓	✗	74.78%	+2.48%
✓	✓	✓	✓	75.10%	+2.80%

Table 4. Combinations of sub-modules in STAA with TSRD.

we conduct experiments on the DVS128 Gesture and CIFAR10-DVS datasets. Unlike static datasets, these neuromorphic datasets contain temporal event streams, requiring models to efficiently process spiking sequences. As shown in Tab. 3, our model achieves state-of-the-art or superior performance with a smaller network and fewer timesteps. Specifically, STAA-SNNs achieve 82.10% accuracy on CIFAR10-DVS and 98.61% on the DVS128 Gesture dataset, demonstrating their effectiveness in event-based action recognition tasks.

4.3. Ablation Study

All ablation studies with T=4 using ResNet-20 as baseline.

Position encoding placement. To optimize PE in SNNs, we evaluate its best placement within the network. Given the structural complexity of SNNs compared to CNNs, identifying the most effective integration point is crucial. As shown in Fig. 2, we evaluate four placement strategies (b)-(e) alongside a baseline SNN without PE (a), ensuring all other conditions remain constant. The CIFAR-10 accuracies for (b) to (e) are 94.08%, 94.29%, 94.27%, and 93.96%, respectively. The highest performance occurs when PE is placed at the pre-aggregation spatial input port (c). Although (b) and (d) yield similar results, (d) introduces additional parameters without significant gains. Thus, placing PE at the pre-aggregation spatial input port achieves the best balance between accuracy and efficiency.

Evaluation of position encoding methods. We next evaluate different encoding methods to determine the most effective approach. The results on CIFAR-10 indicate that learnable PE achieves an accuracy of 94.99%, outperforming fixed PE, which attains 94.42%. This 0.57% improvement highlights the advantage of adaptive position encoding, as it enables the network to dynamically adjust spatial representations, leading to better feature alignment and overall performance.

Impact of dropout probability β in TSRD. To accelerate training and enhance generalization in TSRD, it is essential to determine an optimal dropout probability β. Excessive dropout may lead to network collapse by discarding too much information, while insufficient dropout may limit its regularization effect. As shown in Fig. 3, by incrementally increasing β from 0.1 to 1.0, the model achieves relatively high performance when $\beta \leq 0.3$, and achieving optimal at 0.1. These results confirm that excessive dropout removes essential information, degrading performance, while a balanced dropout probability effectively suppresses noise in deeper timesteps while preserving critical information in shallower steps. This enables the network to navigate critical learning stages more effectively, enhancing both stability and robustness.

Combinations of sub-modules in STAA with TSRD. To evaluate the contribution of each sub-module in STAA, as well as the impact of TSRD, we conduct an ablation study by progressively integrating these components into the SNN. The results on CIFAR-100 are presented in Tab. 4. Notably, the GC block alone improves accuracy by 0.92%, while the SA block contributes an additional 1.14%, highlighting the importance of spatio-temporal attention and feature aggregation reconstruction in enhancing SNN performance. When all components, including TSRD, are incorporated, the accuracy reaches 75.10%, demonstrating a cumulative improvement of 2.80% over the baseline.

Impact of Intermediate Dimension Scaling Coefficients in GC. The GC module incorporates multiple 1×1 convolutions for feature transformation. Retaining the original channel size C for the last two convolutions layers would significantly increase computational overhead. To maintain the module's efficiency while preserving performance, a scaling coefficient r is introduced within the GC to compress feature dimensions. By applying this scaling mechanism, the parameter count is reduced from $(C \cdot C)$ to $(2C \cdot C/r)$, making the module more computationally feasible. We conduct experiments to assess the impact of different r values, as shown in Fig. 5. The results indicate that the model achieves optimal performance when $r = 4$, effectively balancing computational efficiency and accuracy.

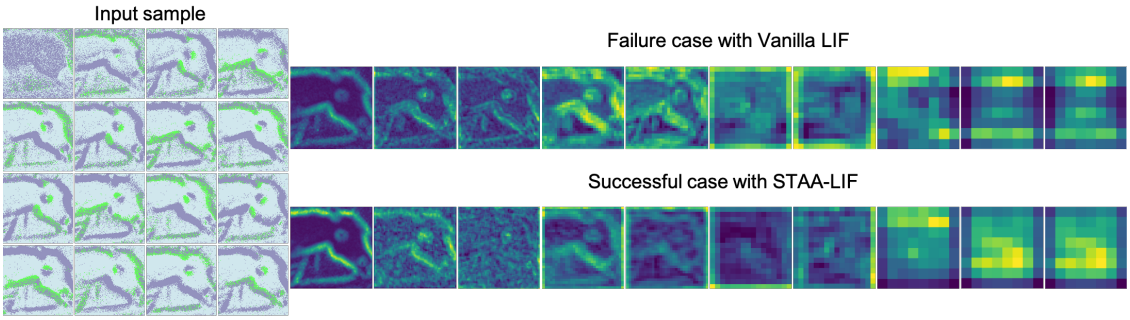

Figure 4. Visualization on CIFAR10-DVS. Ten layers from VGG-13 in a shallow to deep manner.

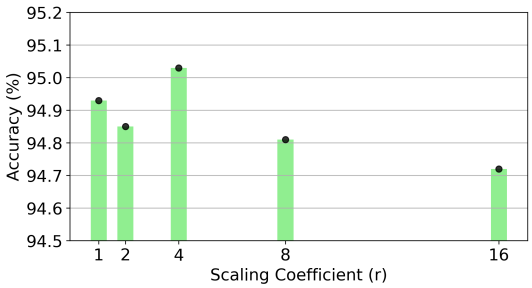

Figure 5. Impact of different scaling coefficients r for intermediate feature dimensions in the GC module on the CIFAR-10 dataset.

4.4. Visualization and Analysis

We aggregate 4D (T, C, H, W) spiking maps into 2D (H, W) and apply Grad-CAM to visualize feature extraction at different depths. As shown in Fig. 4, the left displays a sample from the horse class in CIFAR10-DVS, while the right compares vanilla LIF and STAA-LIF feature maps. STAA-LIF focuses more effectively on target-relevant features, capturing finer details like the horse's mouth, which the vanilla LIF model misses. This highlights the advantage of spatio-temporal attention in enhancing feature selectivity.

4.5. Analysis of Computation Efficiency

We evaluate the single-image inference energy cost using 45-nm technology. ANNs rely on multiplication-and-accumulation (MAC) operations, which consume 4.6 pJ per operation, whereas SNNs primarily use accumulation (AC) operations at a significantly lower cost of 0.9 pJ per operation, making SNNs inherently more energy-efficient [31, 49]. Following [11], we assess the efficiency of ResNet-20, which requires 0.10 billion ACs, 0.06 billion MACs, 0.87 billion FLOPs, and consists of 12.69 million parameters. Under these conditions, inference on CIFAR-100 (with a 32×32 resolution and T=4) results in an energy consumption of 0.366 mJ, demonstrating the energy-efficient nature of ResNet-20. Additional computational efficiency experiments are provided in the Supplementary Material.

5. Conclusion

In this paper, we introduce STAA-SNN, a spatio-temporal attention-based feature aggregator, together with the TSRD training strategy. STAA-SNN leverages the inherent ability of SNNs to process spatio-temporal information, thereby enhancing network performance while reducing the number of required timesteps. Our approach achieves state-of-the-art results across multiple benchmarks. STAA-SNN is designed as a plug-and-play module that seamlessly integrates into convolution-based SNN architectures. It is composed of several synergistic modules that collectively enhance its capabilities. The GC module simulates the brain's ability to process information across a broad spatial range, involving multiple brain regions in coordination. By focusing on spatial feature relevance, it enables the model to capture global structures and key stimuli in its environment. The PE module enhances temporal encoding by incorporating time-step information into feature representations, mirroring the brain's ability to integrate temporal and spatial cues for improved understanding of time-dependent relationships. The SA module mimics the brain's dynamic attention mechanisms, adjusting feature importance across time steps by adaptively enhancing or suppressing information. This prioritization of critical temporal features strengthens decision-making and overall network efficiency. Together, these modules form a robust spatio-temporal feature enhancement system that improves cognitive processing, robustness, and overall model performance. Despite these promising results, future research should focus on reducing computational energy consumption to further improve efficiency, particularly for resource-constrained applications.

6. Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC) (62476035, 62206037, 62276230, 62271361), and Natural Science Foundation of Zhejiang Province (LDT23F02023F02), and State Key Laboratory (SKL) of Biobased Transportation Fuel Technology.

References

[1] Larry F Abbott. Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain research bulletin*, 50(5-6): 303–304, 1999. 1

[2] Filipp Akopyan, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam, Yutaka Nakamura, Pallab Datta, Gi-Joon Nam, et al. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. *IEEE transactions on computer-aided design of integrated circuits and systems*, 34 (10):1537–1557, 2015. 1

[3] Pepe Alcamí and Alberto E Pereda. Beyond plasticity: the dynamic impact of electrical synapses on neural circuits. *Nature Reviews Neuroscience*, 20(5):253–271, 2019. 1

[4] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017. 5, 1

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lui, and Cordelia Schmid. Vivit: A video vision transformer. 2021. 1

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? 2021. 1

[7] Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of neuroscience*, 18(24):10464–10472, 1998. 2

[8] Wuque Cai, Hongze Sun, Rui Liu, Yan Cui, Jun Wang, Yang Xia, Dezhong Yao, and Daqing Guo. A Spatial–Channel–Temporal-Fused Attention for Spiking Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2023. Conference Name: IEEE Transactions on Neural Networks and Learning Systems. 2

[9] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 4

[10] Biswadeep Chakraborty, Xueyuan She, and Saibal Mukhopadhyay. A Fully Spiking Hybrid Neural Network for Energy-Efficient Object Detection. *IEEE Transactions on Image Processing*, 30:9014–9029, 2021. Conference Name: IEEE Transactions on Image Processing. 3

[11] Guangyao Chen, Peixi Peng, Guoqi Li, and Yonghong Tian. Training Full Spike Neural Networks via Auxiliary Accumulation Pathway, 2023. arXiv:2301.11929 [cs]. 8

[12] Dennis V Christensen, Regina Dittmann, Bernabe Linares-Barranco, Abu Sebastian, Manuel Le Gallo, Andrea Redaelli, Stefan Slesazeck, Thomas Mikolajick, Sabina Spiga, Stephan Menzel, et al. 2022 roadmap on neuromorphic computing and engineering. *Neuromorphic Computing and Engineering*, 2(2):022501, 2022. 1

[13] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1):82–99, 2018. 1

[14] Haoyu Deng, Ruijie Zhu, Xuerui Qiu, Yule Duan, Malu Zhang, and Liang-Jian Deng. Tensor decomposition based attention module for spiking neural networks. *Knowledge-Based Systems*, 295:111780, 2024. 5

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. ISSN: 1063-6919. 5

[16] Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. *arXiv preprint arXiv:2202.11946*, 2022. 5, 6

[17] 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. 1, 2, 6

[18] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2661–2671, 2021. 6

[19] Steve B Furber, Francesco Galluppi, Steve Temple, and Luis A Plana. The spinnaker project. *Proceedings of the IEEE*, 102(5):652–665, 2014. 1

[20] Banglei Guan and Ji Zhao. Relative pose estimation for multi-camera systems from point correspondences with scale ratio. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 5036–5044, New York, NY, USA, 2022. Association for Computing Machinery. 1

[21] Banglei Guan, Ji Zhao, Zhang Li, Fang Sun, and Friedrich Fraundorfer. Minimal solutions for relative pose with a single affine correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[22] Yufei Guo, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Yinglei Wang, Xuhui Huang, and Zhe Ma. Im-loss: information maximization loss for spiking neural networks. *Advances in Neural Information Processing Systems*, 35:156–166, 2022. 2, 5, 6

[23] Yufei Guo, Xinyi Tong, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Zhe Ma, and Xuhui Huang. RecDis-SNN: Rectifying Membrane Potential Distribution for Directly Training Spiking Neural Networks. In *Proceedings of CVPR*, pages 326–335, New Orleans, LA, USA, 2022. IEEE. 6

[24] Yufei Guo, Liwen Zhang, Yuanpei Chen, Xinyi Tong, Xiaode Liu, YingLei Wang, Xuhui Huang, and Zhe Ma. Real spike: Learning real-valued spikes for spiking neural networks. In *European Conference on Computer Vision*, pages 52–68. Springer, 2022. 6

[25] Yufei Guo, Xiaode Liu, Yuanpei Chen, Liwen Zhang, Weihang Peng, Yuhan Zhang, Xuhui Huang, and Zhe Ma. RMP-

Loss: Regularizing Membrane Potential Distribution for Spiking Neural Networks, 2023. arXiv:2308.06787 [cs]. 6

[26] Yufei Guo, Yuhan Zhang, Yuanpei Chen, Weihang Peng, Xiaode Liu, Liwen Zhang, Xuhui Huang, and Zhe Ma. Membrane potential batch normalization for spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19420–19430, 2023. 2, 5, 6

[27] Bing Han and Kaushik Roy. Deep spiking neural network: Energy efficiency through time based coding. In *European Conference on Computer Vision*, pages 388–404. Springer, 2020. 2

[28] Bing Han, Gopalakrishnan Srinivasan, and Kaushik Roy. Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13558–13567, 2020. 2

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[30] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005. 2

[31] Mark Horowitz. 1.1 Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14, 2014. ISSN: 2376-8606. 8, 3

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[33] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):2011–2023, 2020. 2

[34] Yifan Hu, Yujie Wu, Lei Deng, and Guoqi Li. Advancing residual learning towards powerful deep spiking neural networks. *arXiv e-prints*, pages arXiv–2112, 2021. 1, 2

[35] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010. 5, 1

[36] Souvik Kundu, Gourav Datta, Massoud Pedram, and Peter A Beerel. Spike-thrift: Towards energy-efficient deep spiking neural networks by limiting spiking activity via attention-guided compression. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3953–3962, 2021. 2

[37] Gabriele Lagani, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. Spiking neural networks and bio-inspired supervised deep learning: A survey. *arXiv preprint arXiv:2307.16235*, 2023. 1

[38] Chankyu Lee, Syed Shakib Sarwar, Priyadarshini Panda, Gopalakrishnan Srinivasan, and Kaushik Roy. Enabling spike-based backpropagation for training deep neural network architectures. *Frontiers in neuroscience*, page 119, 2020. 1, 2

[39] Taihang Lei, Banglei Guan, Minzu Liang, Zibin Liu, Jianbin Liu, Yang Shang, and Qifeng Yu. Motion measurements of explosive shock waves based on an event camera. *Optics Express*, 32(9):15390–15409, 2024. 5

[40] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017. 5, 1

[41] Yudong Li, Yunlin Lei, and Xu Yang. Spikeformer: A novel architecture for training high-performance low-latency spiking neural network. *arXiv preprint arXiv:2211.10686*, 2022. 3, 6

[42] Shuang Lian, Jiangrong Shen, Qianhui Liu, Ziming Wang, Rui Yan, and Huajin Tang. Learnable surrogate gradient for direct training spiking neural networks. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 3002–3010, 2023. 5, 6

[43] Shuang Lian, Jiangrong Shen, Ziming Wang, and Huajin Tang. IM-LIF: Improved Neuronal Dynamics With Attention Mechanism for Direct Training Deep Spiking Neural Network. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–11, 2024. Conference Name: IEEE Transactions on Emerging Topics in Computational Intelligence. 2, 5, 6

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. 2021. 1

[45] Zibin Liu, Banglei Guan, Yang Shang, Qifeng Yu, and Laurent Kneip. Line-based 6-dof object pose estimation and tracking with an event camera. *IEEE Transactions on Image Processing*, 2024. 5

[46] Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36 (6):51–63, 2019. 2

[47] Kleanthis C Neokleous, Marios N Avraamides, Costas K Neocleous, and Christos N Schizas. Selective attention and consciousness: investigating their relation through computational modelling. *Cognitive Computation*, 3:321–331, 2011. 2

[48] Garrick Orchard, E Paxon Frady, Daniel Ben Dayan Rubin, Sophia Sanborn, Sumit Bam Shrestha, Friedrich T Sommer, and Mike Davies. Efficient neuromorphic signal processing with loihi 2. In *2021 IEEE Workshop on Signal Processing Systems (SiPS)*, pages 254–259. IEEE, 2021. 1

[49] 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, 2015. Publisher: Frontiers. 8, 3

[50] Nitin Rathi and Kaushik Roy. Diet-snn: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 5

[51] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019. 1

[52] Chen She and Laiyun Qing. Spikeformer: Image reconstruction from the sequence of spike camera based on transformer. In *Proceedings of the 2022 5th International Conference on Image and Graphics Processing*, pages 72–78, 2022. 3

[53] Gina G Turrigiano and Sacha B Nelson. Homeostatic plasticity in the developing nervous system. *Nature reviews neuroscience*, 5(2):97–107, 2004. 1

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2017. 2

[55] Bingsen Wang, Jian Cao, Jue Chen, Shuo Feng, and Yuan Wang. A new ann-snn conversion method with high accuracy, low latency and good robustness. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 3067–3075, 2023. 2

[56] Yuchen Wang, Kexin Shi, Chengzhuo Lu, Yuguo Liu, Malu Zhang, and Hong Qu. Spatial-Temporal Self-Attention for Asynchronous Spiking Neural Networks. In *Proceedings of IJCAI*, pages 3085–3093, Macau, SAR China, 2023. 6

[57] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018. 1, 3

[58] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1311–1318, 2019. 2, 3

[59] Yujie Wu, Rong Zhao, Jun Zhu, Feng Chen, Mingkun Xu, Guoqi Li, Sen Song, Lei Deng, Guanrui Wang, Hao Zheng, et al. Brain-inspired global-local learning incorporated with neuromorphic computing. *Nature Communications*, 13(1): 65, 2022. 1

[60] Qi Xu, Yaxin Li, Xuanye Fang, Jiangrong Shen, Jian K. Liu, Huajin Tang, and Gang Pan. Biologically inspired structure learning with reverse knowledge distillation for spiking neural networks, 2023. arXiv:2304.09500 [cs]. 2

[61] Qi Xu, Jie Deng, Jiangrong Shen, Biwu Chen, Huajin Tang, and Gang Pan. Hybrid Spiking Vision Transformer for Event-Based Object Detection, 2024. 2

[62] Qi Xu, Xuanye Fang, Yaxin Li, Jiangrong Shen, De Ma, Yi Xu, and Gang Pan. Rsnn: Recurrent spiking neural networks for dynamic spatial-temporal information processing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10602–10610, 2024. 2

[63] Kashu Yamazaki, Viet-Khoa Vo-Ho, Darshan Bulsara, and Ngan Le. Spiking neural networks and their applications: A review. *Brain Sciences*, 12(7):863, 2022. 1

[64] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10221–10230, 2021. 2

[65] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-Wise Attention Spiking Neural Networks for Event Streams Classification. In *Proceedings of ICCV*, pages 10221–10230, 2021. 6

[66] Man Yao, Guangshe Zhao, Hengyu Zhang, Yifan Hu, Lei Deng, Yonghong Tian, Bo Xu, and Guoqi Li. Attention spiking neural networks. *arXiv preprint arXiv:2209.13929*, 2022. 2, 6

[67] Man Yao, Jiakui Hu, Guangshe Zhao, Yaoyuan Wang, Ziyang Zhang, Bo Xu, and Guoqi Li. Inherent Redundancy in Spiking Neural Networks. In *Proceedings of the ICCV*, pages 16924–16934, 2023. 6

[68] Man Yao, Guangshe Zhao, Hengyu Zhang, Yifan Hu, Lei Deng, Yonghong Tian, Bo Xu, and Guoqi Li. Attention Spiking Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9393–9410, 2023. 3

[69] 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. 5, 6

[70] Xingting Yao, Fanrong Li, Zitao Mo, and Jian Cheng. Glif: A unified gated leaky integrate-and-fire neuron for spiking neural networks. *Advances in Neural Information Processing Systems*, 35:32160–32171, 2022. 5, 6

[71] Chengting Yu, Zheming Gu, Da Li, Gaoang Wang, Aili Wang, and Erping Li. STSC-SNN: Spatio-Temporal Synaptic Connection with temporal convolution and attention for spiking neural networks. *Frontiers in Neuroscience*, 16, 2022. 2

[72] Malu Zhang, Jibin Wu, Yansong Chua, Xiaoling Luo, Zihan Pan, Dan Liu, and Haizhou Li. Mpd-al: an efficient membrane potential driven aggregate-label learning algorithm for spiking neurons. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1327–1334, 2019. 2

[73] Hanle Zheng, Yujie Wu, Lei Deng, Yifan Hu, and Guoqi Li. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11062–11070, 2021. 1, 6

[74] Chenlin Zhou, Liutao Yu, Zhaokun Zhou, Zhengyu Ma, Han Zhang, Huihui Zhou, and Yonghong Tian. Spikingformer: Spike-driven residual learning for transformer-based spiking neural network. *arXiv preprint arXiv:2304.11954*, 2023. 5

[75] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. *arXiv preprint arXiv:2209.15425*, 2022. 5

[76] Rui-Jie Zhu, Malu Zhang, Qihang Zhao, Haoyu Deng, Yule Duan, and Liang-Jian Deng. Tcja-snn: Temporal-channel joint attention for spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2