Advancing Spiking Neural Networks Towards Multiscale Spatiotemporal Interaction Learning

Yimeng Shan[1,2], Malu Zhang[2], Rui-jie Zhu[3], Xuerui Qiu[2], Jason K. Eshraghian[3], Haicheng Qu[1]*

[1]Liaoning Technical University, China
[2]University of Electronic Science and Technology of China, China
[3]University of California, Santa Cruz, USA
yimengshan2001@gmail.com, quhaicheng@lntu.edu.cn

Abstract

Recent advancements in neuroscience research have propelled the development of Spiking Neural Networks (SNNs), which not only have the potential to further advance neuroscience research but also serve as an energy-efficient alternative to Artificial Neural Networks (ANNs) due to their spike-driven characteristics. However, previous studies often overlooked the multiscale information and its spatiotemporal correlation between event data, leading SNN models to approximate each frame of input events as static images. We hypothesize that this oversimplification significantly contributes to the performance gap between SNNs and traditional ANNs. To address this issue, we have designed a Spiking Multiscale Attention (SMA) module that captures multiscale spatiotemporal interaction information. Furthermore, we developed a regularization method named Attention ZoneOut (AZO), which utilizes spatiotemporal attention weights to reduce the model's generalization error through pseudo-ensemble training. Our approach has achieved state-of-the-art results on mainstream neuromorphic datasets. Additionally, we have reached a performance of 77.1% on the Imagenet-1K dataset using a 104-layer ResNet architecture enhanced with SMA and AZO. This achievement confirms the state-of-the-art performance of SNNs with non-transformer architectures and underscores the effectiveness of our method in bridging the performance gap between SNN models and traditional ANN models.

Code — https://github.com/YmShan/SMA-AZO

Introduction

The biological brain has long served as a rich source of inspiration for the development of neural networks, with Artificial Neural Networks (ANNs) achieving impressive results (Szegedy et al. 2015; Redmon et al. 2016) by mimicking the visual cortex's hierarchical structure. However, the increasing energy consumption of ANNs has emerged as a critical limitation. Spiking Neural Networks (SNNs) that utilize binary spiking signals, offer inherent low-power characteristics due to their non-continuous activation (Maass 1997) and spike-driven properties (Roy, Jaiswal, and Panda 2019). Neuromorphic chips (Davies et al. 2018; Merolla et al. 2014) are expected to accelerate SNN adoption, with the primary goal being to achieve brain-inspired intelligence by combining insights from high-performance deep learning and biological brain mechanisms.

Initially, researchers in Spiking Neural Networks (SNNs) faced challenges with training algorithms, leading to methods such as STDP (Tao et al. 2023), ANN2SNN (Deng and Gu 2021), and STBP (Wu et al. 2018). Subsequent studies incorporated deep learning elements (e.g., VGG (Sengupta et al. 2019) and ResNet (Fang et al. 2021a; Hu et al. 2024) architectures), neuroscience-inspired features (e.g., attention mechanisms (Yao et al. 2023; Zhu et al. 2022) and biologically plausible neuron models (Wang et al. 2024; Zhang et al. 2024; Yao et al. 2024c)), and brain-inspired learning algorithms (Wei et al. 2023), all aimed at enhancing both performance and neuromorphic characteristics.

However, extant research has insufficiently addressed the heterogeneity in resolution and structural composition among sample features within mainstream datasets. Most studies have focused on developing comprehensive network architectures rather than leveraging multiscale features inherent in complex spatiotemporally correlated data. Furthermore, with the notable exception of Temporal-Channel Joint Attention (TCJA) (Zhu et al. 2022), spatiotemporal correlations have been largely overlooked. We posit that this oversight has resulted in contemporary Spiking Neural Networks (SNNs) relinquishing certain brain-inspired characteristics, potentially compromising their biological plausibility and efficacy.

Therefore, we propose a Spiking Multiscale Attention (SMA) module to introduce multiscale representation learning into the SNN domain, while simultaneously leveraging spatiotemporal correlations to compute attention weights to address this issue. This method enhances the model's ability to extract multiscale features while altering the learning pattern of SNNs, enabling SNN models to better balance local and global features. It is worth noting that we attribute the performance gap between SNNs and ANNs to the insufficient utilization of spatiotemporal correlation information by SNN models. To further mitigate this limitation, we devised an Attention Zoneout (AZO) regularization method. This method improves model generalization by replacing information at spatiotemporal weak points of hidden units with information from previous time step. Unlike dropout,

*Corresponding author

the AZO regularization method involves substitution rather than deletion, facilitating smoother propagation of gradient and state information through time.

The primary contributions of this paper are summarized as follows:

- We observe that previous works have overlooked multiscale information and spatiotemporal correlation, which leads to the loss of brain-inspired features in SNN models, consequently resulting in erroneous learning patterns.
- We extend the Squeeze-and-Excitation (SE module (Hu, Shen, and Sun 2018)) to multiple scales and propose the SMA module based on it to introduce multiscale representation learning into SNNs, while utilizing spatiotemporal correlation information to compute attention weights for balancing local and global information, thereby steering SNN models towards a more neuromorphic learning pattern. To the best of our knowledge, this is the first attempt to integrate multiscale representation learning into SNNs.
- To further leverage spatiotemporal correlation information, we introduce an AZO regularization method leveraging spatiotemporal attention weights. This method employs the previous hidden unit value as noise to replace irrelevant information, thereby training a pseudo ensemble to enhance the robustness and generalization of the model.
- We demonstrate the effectiveness of our proposed approach by achieving state-of-the-art performance on three mainstream neuromorphic datasets and Imagenet-1K. Through comprehensive visualization analysis, we provide evidence that simultaneously leveraging multiscale information and spatiotemporal correlation can indeed induce SNNs to adopt a more brain-inspired learning pattern.

Related Works

Attention Mechanism has become crucial in enhancing deep learning model performance. In SNNs, Yao et al. introduced Time-Attention (TA) (Yao et al. 2021) and Multi-dimensional Attention (MA) modules (Yao et al. 2023), focusing on the importance of temporal and multi-dimensional features, respectively. Zhu et al. proposed the TCJA module (Zhu et al. 2022) for efficient time and channel attention. Shan et al. integrated discrete spiking signals into attention-based decision-making (Shan et al. 2023). Addressing the need for lightweight design and multiscale spatiotemporal correlation handling, we developed the MSE module and its derivative SMA module, which efficiently transform the model's learning pattern.

Multiscale Representation Learning plays a pivotal role in diverse computer vision tasks (Han et al. 2018; Sharon, Brandt, and Basri 2000). This approach addresses the challenges posed by the heterogeneity of shapes and resolutions inherent in natural objects (Jiao et al. 2021). To improve deep learning frameworks, researchers have incorporated various multiscale techniques: multiscale convolution structures (Li et al. 2019), pyramid architectures (Zhang et al. 2016), multiscale loss functions (Lin et al. 2017), and multiscale attention mechanisms (Zhang, Zheng, and Liu 2021). This study extends the application of multiscale representation learning to SNNs by introducing a novel multiscale attention module, thereby broadening the scope of this approach in neuromorphic computing.

Regularization balances bias and variance to reduce generalization error. Data regularization methods (e.g., Cutout (DeVries and Taylor 2017), Mixup (Zhang et al. 2017), CutMix (Yun et al. 2019)) improve model robustness through input transformations. Structural regularization, like Dropout (Wang and Manning 2013) and its variants (e.g., Zoneout (Krueger et al. 2016), Dropblock (Ghiasi, Lin, and Le 2018)) improves performance by selectively retaining activation values. Recent research explores optimal dropout locations, such as Maxdropout (do Santos et al. 2021) and Autodropout (Pham and Le 2021). Our proposed AZO regularization method utilizes SMA attention weights for selection, replacing irrelevant information with previous timestep hidden unit values as noise, training a pseudo-ensemble to enhance robustness and generalizability.

Method

Leaky Integrate-and-Fire Neuron Model

The Leaky Integrate-and-Fire (LIF) neuron (Maass 1997) stands as one of the predominant neuron models within SNNs, esteemed for its balanced performance and adherence to biological principles. It is uniformly adopted in this work. Within neural networks, neurons serve as fundamental computational units. Upon receiving transmitted spiking signals, LIF neurons initiate an integration process. When the membrane potential reaches a threshold, neurons emit spikes and reset their membrane potentials. This dynamic process is encapsulated by the subthreshold dynamics model (Roy, Jaiswal, and Panda 2019)

$$\tau \frac{d\boldsymbol{V}(t)}{dt} = -(\boldsymbol{V}(t) - \boldsymbol{V}_{reset}) + \boldsymbol{I}(t), \tag{1}$$

where τ represents a time constant, $\boldsymbol{V}(t)$ denotes the membrane potential of the postsynaptic neuron, and $\boldsymbol{I}(t)$ signifies the input gathered from presynaptic neurons. Additionally, $\boldsymbol{V}_{\text{reset}}$ denotes the reset potential, which is established subsequent to the activation of the output spiking. To facilitate training and description, we adopt the displayed iteration version of the subthreshold dynamics model (Neftci, Mostafa, and Zenke 2019).

$$\boldsymbol{U}_t^n = \boldsymbol{H}_{t-1}^n + \frac{1}{\tau}(\boldsymbol{I}_{t-1}^n - (\boldsymbol{H}_{t-1}^n - \boldsymbol{U}_{reset})), \tag{2}$$

$$\boldsymbol{S}_t^n = \Theta(\boldsymbol{U}_t^n - \boldsymbol{U}_{threshold}), \tag{3}$$

$$\boldsymbol{H}_t^n = \boldsymbol{U}_t^n(1 - \boldsymbol{S}_t^n), \tag{4}$$

at each layer n and time step t, the membrane potential $\boldsymbol{U}$ of a neuron is denoted as $\boldsymbol{U}_{\text{t}}^{\text{n}}$. The parameter τ signifies a time constant, and $\boldsymbol{S}$ represents a binary spiking tensor. $\boldsymbol{I}$ denotes the neuron's input, while Θ represents the Heaviside step function. $\boldsymbol{H}$ symbolizes the hidden state, $\boldsymbol{U}_{\text{reset}}$ refers to the reset potential of the neuron following a spike, and $\boldsymbol{U}_{\text{threshold}}$ indicates the discharge threshold of the neuron.

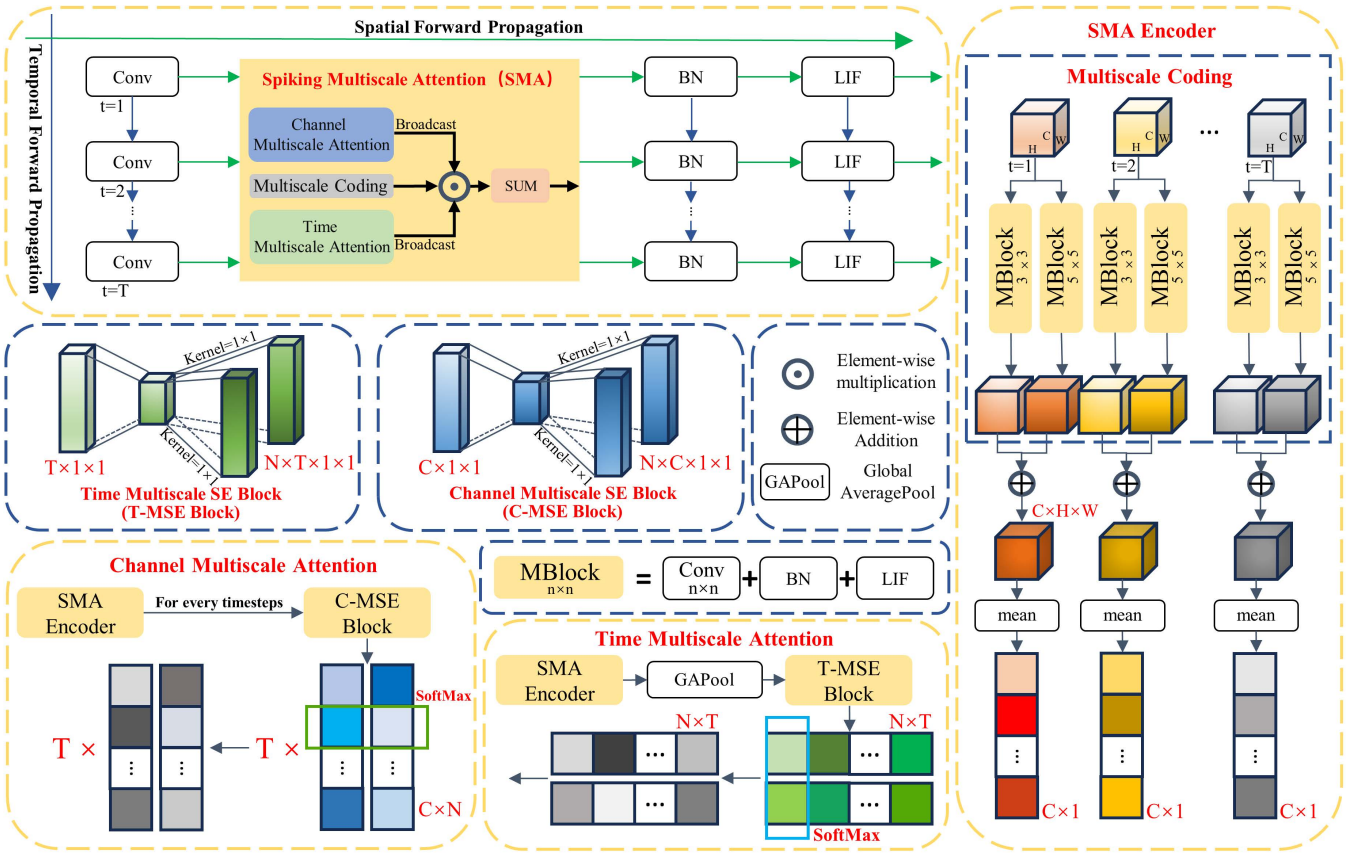

Figure 1: The overview of Spiking Multiscale Attention (SMA) module. In the figure, the schematic diagram of the encoder is shown on the right side, the schematic diagram of the decoder is displayed in the lower-left corner, and the schematic diagram of the Multiscale SE (MSE) block module is positioned in the center.

Spiking Multiscale Attention for SNNs

The overall structure of the SMA module is illustrated in Fig. 1. Its primary objective is to enhance the model's ability to extract target features with diverse resolutions, shapes, and states by utilizing spatiotemporal correlation information. This approach helps balance the significance of global and local features, thereby transforming the model's learning pattern. The temporal and spatial information inherent in event data can often be intertwined, posing additional challenges in extracting multiscale features. Thus, instead of directly conducting multiscale feature extraction within the encoder, we opt to perform multiscale coding beforehand to enhance feature representation. The encoder's architecture is delineated in the right portion of Fig. 1. Given an input event sequence $\boldsymbol{X} = [\cdots, \boldsymbol{X}_t, \cdots] \in \mathbb{R}^{T\times C\times H\times W}$, the encoder of the SMA module across $\boldsymbol{N}$ scales can be represented as

$$E(x,k) = \delta(BN(\text{Conv2}d(k,x))), \tag{5}$$

$$\boldsymbol{M}_t^n = E(\boldsymbol{X}_t, \boldsymbol{K}_t^n), \quad n \in \boldsymbol{N}, t \in \boldsymbol{T}, \boldsymbol{M}_t^n \in \mathbb{R}^{C\times H\times W} \tag{6}$$

$$\boldsymbol{Y}_t = \sum_{n=1}^{N} \frac{\boldsymbol{M}_t^n}{\boldsymbol{N}}, \quad \boldsymbol{Y}_t \in \mathbb{R}^{C\times H\times W} \tag{7}$$

where $\boldsymbol{E}(x,k)$ represents the multiscale coding process and δ denotes the $ReLU$ function. However, when analyzing the mechanism of SMA, we substitute δ with LIF neurons(we have conducted ample experiments in Sec. B.3 of the Supplementary Material demonstrating that LIF does not notably compromise model accuracy compared to $ReLU$). $\text{Conv2}d(x,k)$ indicates that the 2D convolution operation is performed on $\boldsymbol{X}$ using the convolution kernel $\boldsymbol{K}$, and $\boldsymbol{M}_t = [\cdots, \boldsymbol{M}_t^n, \cdots] \in \mathbb{R}^{N\times C\times H\times W}$ represents the multiscale coding result of the t-th frame of the input event.

The decoder structure is illustrated in the lower-left portion of Fig. 1. It is responsible for processing the multiscale data $\boldsymbol{Y} = [\cdots, \boldsymbol{Y}_t, \cdots] \in \mathbb{R}^{T\times C\times H\times W}$, enriched with characteristics output from the encoder. Specifically, it calculates the attention weights for the temporal and channel dimensions as follows:

$$f_\alpha^n(\boldsymbol{E}) = \text{Conv2}d(\boldsymbol{K}_{E,\alpha}^n, \delta(\text{Conv2}d(\boldsymbol{K}_{S,\alpha}, \boldsymbol{E}))), \tag{8}$$

$$f_\beta^n(\boldsymbol{E}) = \text{Conv2}d(\boldsymbol{K}_{E,\beta}^n, \delta(\text{Conv2}d(\boldsymbol{K}_{S,\beta}, \boldsymbol{E}))), \tag{9}$$

$$\boldsymbol{W}_\alpha = \eta(f_\alpha(Avgpool(\boldsymbol{Y}_t))), \quad \boldsymbol{W}_\alpha \in \mathbb{R}^{N\times T\times 1} \tag{10}$$

$$\boldsymbol{W}_{\beta,t} = \eta(f_\beta(\boldsymbol{Y}_t)), \quad \boldsymbol{W}_{\beta,t} \in \mathbb{R}^{N\times C\times 1} \tag{11}$$

where $f_{\alpha}^{n}(\boldsymbol{E})$ and $f_{\beta}^{n}(\boldsymbol{E})$ respectively describe the roles of the MSE (T-MSE) module in the time dimension and the MSE (C-MSE) module in the channel dimension. $\boldsymbol{K}_{S,\alpha}$ and $\boldsymbol{K}_{S,\beta}$ represent the Squeeze convolution kernel (1×1) in the T-MSE module and C-MSE module, respectively. $\boldsymbol{K}_{E,\alpha}^{n}$ and $\boldsymbol{K}_{E,\beta}^{n}$ represent the Excitation convolution kernel (1×1) in the n-th scale, respectively. η represents the $SoftMax$ operation.

To compute attention weights in the time dimension, we adhere to the conventional methodology: initially conducting global average pooling on the input data, followed by computing attention weights to derive weights sized $T \times 1$. To comprehensively leverage the interplay between multiscale information and spatiotemporal interactions, attention weights are computed for all input channels at each timestep. It is noteworthy that upon completing the calculation, $SoftMax$ operations are performed on the weights within the time and channel dimensions correspondingly, in the scale dimension, to mitigate the adverse effects of large (small) values on the model. Consequently, $\boldsymbol{W}_{\beta} = [\cdots, \boldsymbol{W}_{\beta,t}, \cdots] \in \mathbb{R}^{N \times T \times C \times 1}$ is obtained. Finally, apply the attention weights to the multiscale encoding result $\boldsymbol{M} = [\cdots, \boldsymbol{M}_t, \cdots] \in \mathbb{R}^{N \times T \times C \times H \times W}$ of the input event stream and aggregate them along the scale dimension:

$$\boldsymbol{Z} = Sum(\boldsymbol{M} \times \boldsymbol{W}_{\alpha} \times \boldsymbol{W}_{\beta}). \quad \boldsymbol{Z} \in \mathbb{R}^{T \times C \times H \times W} \tag{12}$$

Attention Zoneout for SMA

Algorithm 1: Attention Zoneout

Input: Output of SMA module decoder : $\boldsymbol{Z}$; Time attention weight : $\boldsymbol{W}_{\alpha}$; Channel attention weight : $\boldsymbol{W}_{\beta}$; Number of timesteps executing AZO : δ_t; Number of channels executing AZO : δ_c

Output: Z after AZO execution : $\boldsymbol{R}$

(The following operations are concurrently executed at all scales)

$\boldsymbol{R} = \boldsymbol{Z}$

Find the indices of the δ_t smallest values in array $\boldsymbol{W}_{\alpha} : \boldsymbol{H} \in \mathbb{N}^{\delta_t \times 1}$

$\boldsymbol{H} = Sort(\boldsymbol{H}, Ascending)$

for *each* i *in* $\boldsymbol{H}$ **do**

  Find the indices of the δ_c smallest values in array $\boldsymbol{W}_{\beta,i} : \boldsymbol{P}_i \in \mathbb{N}^{\delta_c \times 1}$

end

for *each* i *in* $\boldsymbol{H}$ **do**

  for *each* j *in* $\boldsymbol{P}_i$ **do**

    $\boldsymbol{R}[i][j] = \boldsymbol{Z}[i-1][j] \quad if(i \neq 0)$

  end

end

return $\boldsymbol{R}$

Zhu et al. were the first to explore the correlation between temporal and spatial information in event data within the domain of SNNs (Zhu et al. 2022). However, most mainstream SNN methods overlook this spatiotemporal correlation information. To address this, we propose a regularization method called Attention Zoneout (AZO), which utilizes temporal and channel attention weights, as described in Algorithm 1, to leverage these valuable pieces of information.

Similar to Zoneout (Krueger et al. 2016), AZO trains pseudo-ensembles by adding noise to hidden units, thereby enhancing generalization ability. The key difference is that in AZO, the location of noise is determined by attention weights rather than being random. This helps mitigate the impact of irrelevant features while introducing noise, facilitating the network's convergence to the global optimum. Additionally, while Zoneout directly adds noise in the spatial dimension, AZO applies noise in the channel dimension because, at the time of AZO operation, all features have already been aggregated onto the channel dimension, resulting in a spatial dimension size of 1×1. Consequently, each operation applied to the channel dimension is equivalently projected onto the spatial dimension. In Sec. B.2 of the Supplementary Material, we present an ablation study to elucidate the selection process of the hyperparameters δ_t and δ_c in AZO and provide additional experiments to validate its efficacy.

Experiments

Even with the use of spiking coding, the spatiotemporal interaction in static image datasets remains limited. Therefore, we only evaluated the classification performance of our proposed SMA-SNN and SMA-AZO-SNN architectures on three prominent neuromorphic datasets (DVS128 Gesture (Amir et al. 2017), CIFAR10-DVS (Li et al. 2017), and N-Caltech101 (Orchard et al. 2015)) as well as the ImageNet-1K dataset (Deng et al. 2009). Preceding this assessment, we provided a comprehensive exposition of our attention position determination and scale quantity selection process, a crucial aspect we regard as essential for a generic attention module. All network structures and hyperparameter settings utilized in the experiment are detailed in Sec. A of Supplementary Material.

Ablation Study

Building upon the foundation established by prior research, it has been observed that the classification performance of SNN models on the DVS128 Gesture dataset (Amir et al. 2017) can be effectively extrapolated to other prominent datasets (Yao et al. 2021, 2023; Zhu et al. 2022; Fang et al. 2021a). Consequently, our ablation studies will be systematically conducted utilizing the DVS128 Gesture dataset as the primary benchmark.

SMA Position. Incorporating the plug-and-play attention module into mainstream network architectures necessitates careful consideration of its insertion points and quantity. Initially, we address the unique characteristics of the encoding module (the first module of SNNs) and categorize the potential roles of SMA within the network as follows: **T1, absence of SMA; T2, exclusive addition of SMA to the encoding module; T3, integration of SMA throughout**

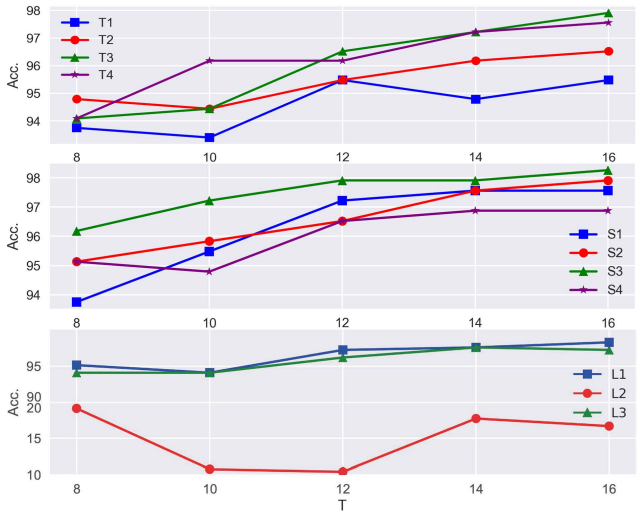

Figure 2: Ablation study of different SMA positions based on DVS128 Gesture. Inspired by previous work (Yao et al. 2021, 2023), we placed SMA behind convolution layer in the first two groups of experiments.

the network except for the encoding module; T4, comprehensive inclusion of SMA across the entire network. As depicted in Fig. 2, T3 consistently outperforms the other configurations across nearly all simulation timesteps, particularly at the prevalent timestep of 16.

Building upon the aforementioned findings, we further subdivide the potential optimal positions for SMA within the network as follows: **S1, incorporating SMA solely in odd-numbered blocks excluding the encoding module; S2, integrating SMA exclusively in even-indexed blocks; S3, embedding SMA only in the initial half of the network, but excluding the coding module; S4, inserting SMA solely in the latter half of the network.** As illustrated in Fig. 2, the S3 configuration demonstrates superior performance compared to other positions across all simulated timesteps. Additionally, we investigated the effect of SMA within the conventional SNN module (Conv-BN-Neuron). The potential insertion points for SMA are: **L1, immediately following the convolution layer; L2, following the BN layer; L3, after the neuron layer.** The experimental findings demonstrate that within the module, the placement of the SMA module at position L1 exhibits a slight advantage over its placement at L3, while positioning it at L2 results in the network's failure to converge. This phenomenon is hypothesized to stem from the SMA module's inability to accurately extract significance from normalized data, consequently leading to the disappearance of the network's gradient.

Scale Quantities. Paying attention to various receptive fields of differing sizes is crucial for SMA. The results in Tab. 1 demonstrate that the SMA module, integrating downsampling and attention mechanisms utilizing filters of sizes 1, 3, 5, and 7 across four distinct scales, yields the most pronounced and consistent effects. Furthermore, this configuration introduces only marginal additional reasoning time,

Scale	Acc(%)	Inference Overhead(s)
No SMA	96.52	18.98
2	96.52(±0.34)	21.99(±0.98)
3	97.21(±0.69)	23.02(±0.07)
4	97.91(±0.34)	27.44(±0.24)
5	97.73(±0.17)	33.72(±0.20)
6	97.56(±0.34)	42.43(±0.06)

Table 1: Various Scale Quantities of SMA. The filter sizes used for sampling at each scale are progressively 1, 3, 5, 7, 9, and 11. The experimental results are derived from ten identical experiments.

thereby facilitating efficient processing for individual events without a significant increase in computational load.

Comparison with the State-of-the-Art

Experimental Results on Imagenet-1K. We adopted the same approach as Yao et al (Yao et al. 2023), simply integrating the SMA module into MS-ResNet (Hu et al. 2024) to evaluate its performance. Tab. 2 presents the experimental results. The integration of SMA alone significantly improved model accuracy, evidenced by enhancements of 5.36% in the 18-layer ResNet and 0.77% in the 34-layer ResNet configurations. Furthermore, the combined application of SMA and AZO resulted in an additional accuracy increase of 1.03% in the 104-layer structure. Comparison to previous studies indicates that our approach achieves state-of-the-art accuracy in traditional convolutional-based SNNs and is also competitive with some models based on the Transformer architecture.

Experimental Results on Mainstream Neuromorphic Datasets. As Tab. 3 illustrates, the integration of SMA into the VGG network achieves SOTA performance on the CIFAR10-DVS and N-Caltech101 datasets. The accuracy is further and steadily improved through the application of the AZO regularization method. Our approach also resulted in a significant 6.1% increase in accuracy on the N-Caltech101 dataset. Moreover, on the DVS128 Gesture dataset, our method achieved equivalent results to those of non-transformer architectures with the same timesteps.

Study of Learning Patterns

SMA Changes SNNs Learning Patterns. The attention heatmaps in Fig. 3(b) show that traditional SNN approaches treat event data frames as static images, overemphasizing global features while neglecting local features and temporal information. In contrast, Fig. 3(c) demonstrates that the SMA module balances local and global feature importance by integrating multiscale and spatiotemporal correlation information. This shift in decision-making focuses on the relative positions and dynamic information of crucial joints, enhancing the utilization of spatiotemporal correlations between event frames.

The N-Caltech101 dataset has more complex features and more noise. A closer examination of Fig. 3 reveals that SNN models lacking SMA tend to allocate unnecessary attention

Work	Architecture	Timestep	Top-1 Acc.(%)
SEW ResNet (Fang et al. 2021a)	SEW-ResNet-34	4	67.04
	SEW-ResNet-50	4	67.78
	SEW-ResNet-101	4	68.76
	SEW-ResNet-152	4	69.26
MS ResNet (Hu et al. 2024)	MS-ResNet-18	6	63.10
	MS-ResNet-34	6	69.42
	MS-ResNet-104	5	74.21
	MS-ResNet-104*	5	76.02
MA-ResNet (Yao et al. 2023)	MA-MS-ResNet-18	1	63.97
	MA-MS-ResNet-34	1	69.15
	MA-MS-ResNet-104*	4	**77.08**
Spiking ResNet (Hu, Tang, and Pan 2021)	ResNet-50	350	72.75
Hybrid training (Rathi et al. 2020)	ResNet-34	250	61.48
TET (Deng et al. 2022)	SEW-ResNet-34	4	68.00
tdBN (Zheng et al. 2021)	Spiking-ResNet-34	6	63.72
Spike-Norm (Sengupta et al. 2019)	ResNet34	2000	65.47
QCFS (Bu et al. 2023)	VGG-16	64	72.85
Fast-SNN (Hu et al. 2023)	VGG-16	7	72.95
ResNet (Hu et al. 2024)	Res-CNN-104	-	76.87
Spikformer (Zhou et al. 2022)	Spiking Transformer-10-512	4	73.68
	Spiking Transformer-8-768	4	74.81
Spike-driven Transformer (Yao et al. 2024b)	Spiking Transformer-10-512	4	74.66
	Spiking Transformer-8-768*	4	**77.07**
Meta-SpikeFormer (Yao et al. 2024a)	Meta-SpikeFormer	4	**77.20**
	Meta-SpikeFormer*	4	**80.00**
SMA-ResNet(Ours)	SMA-MS-ResNet-18	6	**68.46**
	SMA-MS-ResNet-34	6	**70.19**
	SMA-AZO-MS-ResNet-104*	5	**77.05**

Table 2: Evaluation on ImageNet-1K. In inference, the default resolution of the input crops is 224×224. The experimental input crops marked with * are enlarged to 288×288.

Work	Spike-driven	DVS128 Gesture		CIFAR10-DVS		N-Caltech101	
		T	Acc	T	Acc	T	Acc
PLIF (Fang et al. 2021b)	✓	20	97.6	20	74.8	-	-
Spikformer (Zhou et al. 2022)	✗	**16**	**98.3**	16	80.9	-	-
tdBN (Zheng et al. 2021)	✗	40	96.9	10	67.8	-	-
SEW-ResNet (Fang et al. 2021a)	✗	16	97.9	16	74.4	-	-
TA-SNN (Yao et al. 2021)	✗	60	98.6	10	72.0	-	-
HATS (Sironi et al. 2018)	N/A	-	-	N/A	52.4	N/A	64.2
DART (Ramesh et al. 2019)	N/A	-	-	N/A	65.8	N/A	66.8
SALT (Kim and Panda 2021)	✗	-	-	20	67.1	20	55.0
TCJA-SNN (Zhu et al. 2022)	✗	**20**	**99.0**	10	80.7	14	78.5
Spike-driven Transformer (Yao et al. 2024b)	✓	**16**	**99.3**	16	80.0	-	-
SMA-VGG(Ours)	✓	16	98.3	10	**83.1**	14	**83.7**
SMA-AZO-VGG(Ours)	✓	**16**	**98.6**	10	**84.0**	14	**84.6**

Table 3: The comparison between the proposed methods and existing SOTA techniques on three mainstream neuromorphic datasets.

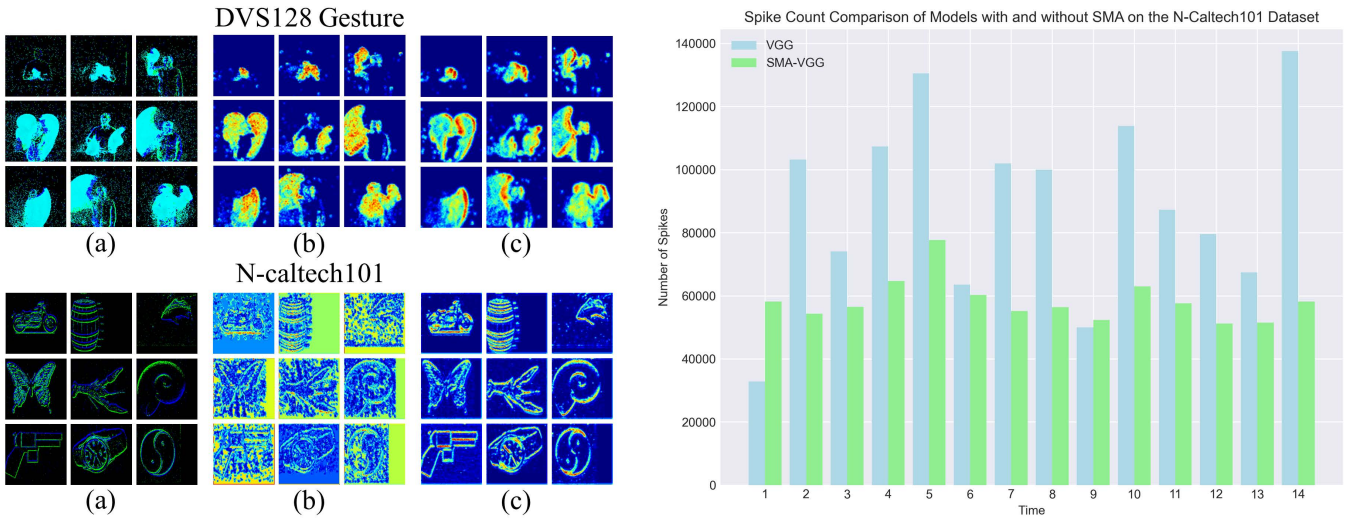

Figure 3: Visualization of typical sample input frames and their attentional heatmaps: (a) shows the input frame; (b) and (c) display attention heat maps based on the Spiking Firing Rate (SFR), where red indicates high and blue shows low spiking activation. Heat map (b) is from the Spiking-VGG8 model and (c) from the SMA-SNN model. All heat maps are from the first convolutional layer of each model, except the coding layer. A spike count comparison is shown on the right.

to the background while neglecting local features. We identify this as the primary reason for the underperformance of traditional SNN models on the N-Caltech101 dataset. Upon integrating SMA, the SMA-SNN model effectively mitigates the issue of background overemphasis observed in traditional SNNs, redirecting attention more precisely towards relevant local features. This enhancement is believed to be the primary factor contributing to the improved performance of the SMA-SNN model on the N-Caltech101 dataset. Additionally, as depicted in the right part of Fig. 3, SMA significantly reduces the SFR of the SNN model by minimizing unnecessary attention to the background, thereby promising greater energy savings for event-driven SNNs.

Study on the Correlation between SMA and Multiscale Information in Samples

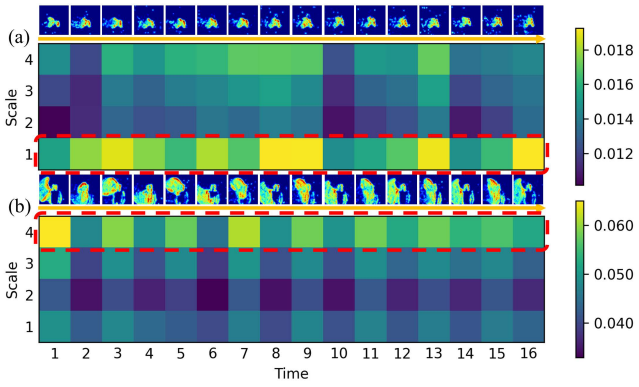

Figure 4: Scale importance between different types of samples.

Focus on Multiscale Features. The filters of sizes 1, 3, 5, and 7 in Fig. 4 are employed for downsampling across the four scales. Research conducted on this figure suggests that in the decision-making process, samples characterized by smaller key features (such as fingers and wrists) exhibit a preference for filters of smaller size, whereas samples with larger key features (such as arms and trajectories) show a predilection for filters of larger size. This observation underscores the intrinsic capability of the SMA module to enhance model performance in datasets with a diverse range of feature sizes.

Conclusion

This study proposes a plug-and-play attention module called SMA that focuses on multiple dimensions. Experiments demonstrate that SMA effectively utilizes both multiscale information and spatiotemporal correlation information to balance the importance of global and local features in the samples, thereby changing the learning patterns of SNNs. We further utilize spatiotemporal correlation information by proposing an AZO regularization method. This method replaces the information at spatiotemporal weak points with the information from the corresponding spatial locations at the previous time step to train pseudo-ensembles, effectively reducing the model's generalization error. SMA and AZO achieve state-of-the-art accuracy on CIFAR10-DVS (84.0%) and N-Caltech101 (84.6%) datasets, while also attaining non-Transformer architecture state-of-the-art accuracy on the ImageNet-1k dataset (77.1%), underscoring their efficacy. We hope that our work can contribute to the research on neuromorphic SNNs, while simultaneously providing inspiration for the development of SNN models that are both higher-performing and more neuromorphic.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grants U20B2063, 62220106008, 62106038, and 42271409; the Scientific Research Foundation of Higher Education Institutions of Liaoning Province under grant LJKMZ20220699; the Sichuan Science and Technology Program under grants 2024NSFTD0034 and 2023YFG0259; and the Open Research Fund of the State Key Laboratory of Brain-Machine Intelligence, Zhejiang University (Grant No. BMI2400020).

References

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

Bu, T.; Fang, W.; Ding, J.; Dai, P.; Yu, Z.; and Huang, T. 2023. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks. *arXiv preprint arXiv:2303.04347*.

Davies, M.; Srinivasa, N.; Lin, T.-H.; Chinya, G.; Cao, Y.; Choday, S. H.; Dimou, G.; Joshi, P.; Imam, N.; Jain, S.; et al. 2018. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1): 82–99.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, S.; and Gu, S. 2021. Optimal conversion of conventional artificial neural networks to spiking neural networks. *arXiv preprint arXiv:2103.00476*.

Deng, S.; Li, Y.; Zhang, S.; and Gu, S. 2022. Temporal efficient training of spiking neural network via gradient reweighting. *arXiv preprint arXiv:2202.11946*.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

do Santos, C. F. G.; Colombo, D.; Roder, M.; and Papa, J. P. 2021. MaxDropout: deep neural network regularization based on maximum output values. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2671–2676. IEEE.

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

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

Ghiasi, G.; Lin, T.-Y.; and Le, Q. V. 2018. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31.

Han, J.; Liang, K.; Zhou, B.; Zhu, X.; Zhao, J.; and Zhao, L. 2018. Infrared small target detection utilizing the multiscale relative local contrast measure. *IEEE Geoscience and Remote Sensing Letters*, 15(4): 612–616.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, Y.; Deng, L.; Wu, Y.; Yao, M.; and Li, G. 2024. Advancing Spiking Neural Networks Toward Deep Residual Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Hu, Y.; Tang, H.; and Pan, G. 2021. Spiking deep residual networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 5200–5205.

Hu, Y.; Zheng, Q.; Jiang, X.; and Pan, G. 2023. Fast-snn: Fast spiking neural network by converting quantized ann. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jiao, L.; Gao, J.; Liu, X.; Liu, F.; Yang, S.; and Hou, B. 2021. Multiscale representation learning for image classification: A survey. *IEEE Transactions on Artificial Intelligence*, 4(1): 23–43.

Kim, Y.; and Panda, P. 2021. Optimizing deeper spiking neural networks for dynamic vision sensing. *Neural Networks*, 144: 686–698.

Krueger, D.; Maharaj, T.; Kramár, J.; Pezeshki, M.; Ballas, N.; Ke, N. R.; Goyal, A.; Bengio, Y.; Courville, A.; and Pal, C. 2016. Zoneout: Regularizing rnns by randomly preserving hidden activations. *arXiv preprint arXiv:1606.01305*.

Li, H.; Liu, H.; Ji, X.; Li, G.; and Shi, L. 2017. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11: 244131.

Li, X.; Wang, W.; Hu, X.; and Yang, J. 2019. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 510–519.

Lin, G.; Milan, A.; Shen, C.; and Reid, I. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1925–1934.

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

Merolla, P. A.; Arthur, J. V.; Alvarez-Icaza, R.; Cassidy, A. S.; Sawada, J.; Akopyan, F.; Jackson, B. L.; Imam, N.; Guo, C.; Nakamura, Y.; et al. 2014. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197): 668–673.

Neftci, E. O.; Mostafa, H.; and Zenke, F. 2019. 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.

Orchard, G.; Jayawant, A.; Cohen, G. K.; and Thakor, N. 2015. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9: 159859.

Pham, H.; and Le, Q. 2021. Autodropout: Learning dropout patterns to regularize deep networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9351–9359.

Ramesh, B.; Yang, H.; Orchard, G.; Le Thi, N. A.; Zhang, S.; and Xiang, C. 2019. Dart: distribution aware retinal transform for event-based cameras. *IEEE transactions on pattern analysis and machine intelligence*, 42(11): 2767–2780.

Rathi, N.; Srinivasan, G.; Panda, P.; and Roy, K. 2020. Enabling deep spiking neural networks with hybrid conversion and spike timing dependent backpropagation. *arXiv preprint arXiv:2005.01807*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Roy, K.; Jaiswal, A.; and Panda, P. 2019. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784): 607–617.

Sengupta, A.; Ye, Y.; Wang, R.; and Roy, K. 2019. Going deeper in spiking neural networks: VGG and residual architectures. *Frontiers in neuroscience*, 13: 425055.

Shan, Y.; Qiu, X.; Zhu, R.-j.; Li, R.; Wang, M.; and Qu, H. 2023. OR Residual Connection Achieving Comparable Accuracy to ADD Residual Connection in Deep Residual Spiking Neural Networks. *arXiv preprint arXiv:2311.06570*.

Sharon, E.; Brandt, A.; and Basri, R. 2000. Fast multiscale image segmentation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, 70–77. IEEE.

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

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Tao, T.; Li, D.; Ma, H.; Li, Y.; Tan, S.; Liu, E.-x.; Schutt-Aine, J.; and Li, E.-P. 2023. A new pre-conditioned STDP rule and its hardware implementation in neuromorphic crossbar array. *Neurocomputing*, 557: 126682.

Wang, S.; and Manning, C. 2013. Fast dropout training. In *international conference on machine learning*, 118–126. PMLR.

Wang, S.; Zhang, D.; Belatreche, A.; Xiao, Y.; Qing, H.; We, W.; Zhang, M.; and Yang, Y. 2024. Ternary spike-based neuromorphic signal processing system. *arXiv preprint arXiv:2407.05310*.

Wei, W.; Zhang, M.; Qu, H.; Belatreche, A.; Zhang, J.; and Chen, H. 2023. Temporal-coded spiking neural networks with dynamic firing threshold: Learning with event-driven backpropagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10552–10562.

Wu, Y.; Deng, L.; Li, G.; and Shi, L. 2018. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12: 323875.

Yao, M.; Gao, H.; Zhao, G.; Wang, D.; Lin, Y.; Yang, Z.; and Li, G. 2021. Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10221–10230.

Yao, M.; Hu, J.; Hu, T.; Xu, Y.; Zhou, Z.; Tian, Y.; Xu, B.; and Li, G. 2024a. Spike-driven Transformer V2: Meta Spiking Neural Network Architecture Inspiring the Design of Next-generation Neuromorphic Chips. *arXiv preprint arXiv:2404.03663*.

Yao, M.; Hu, J.; Zhou, Z.; Yuan, L.; Tian, Y.; Xu, B.; and Li, G. 2024b. Spike-driven-transformer. *Advances in Neural Information Processing Systems*, 36.

Yao, M.; Qiu, X.; Hu, T.; Hu, J.; Chou, Y.; Tian, K.; Liao, J.; Leng, L.; Xu, B.; and Li, G. 2024c. Scaling Spike-driven Transformer with Efficient Spike Firing Approximation Training. *arXiv preprint arXiv:2411.16061*.

Yao, M.; Zhao, G.; Zhang, H.; Hu, Y.; Deng, L.; Tian, Y.; Xu, B.; and Li, G. 2023. Attention spiking neural networks. *IEEE transactions on pattern analysis and machine intelligence*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, D.; Wang, S.; Belatreche, A.; Wei, W.; Xiao, Y.; Zheng, H.; Zhou, Z.; Zhang, M.; and Yang, Y. 2024. Spike-based Neuromorphic Model for Sound Source Localization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.

Zhang, T.; Zheng, X.-Q.; and Liu, M.-X. 2021. Multiscale attention-based LSTM for ship motion prediction. *Ocean Engineering*, 230: 109066.

Zheng, H.; Wu, Y.; Deng, L.; Hu, Y.; and Li, G. 2021. Going deeper with directly-trained larger spiking neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11062–11070.

Zhou, Z.; Zhu, Y.; He, C.; Wang, Y.; Yan, S.; Tian, Y.; and Yuan, L. 2022. Spikformer: When spiking neural network meets transformer. *arXiv preprint arXiv:2209.15425*.

Zhu, R.-J.; Zhao, Q.; Zhang, T.; Deng, H.; Duan, Y.; Zhang, M.; and Deng, L.-J. 2022. Tcja-snn: Temporal-channel joint attention for spiking neural networks. *arXiv preprint arXiv:2206.10177*.