DTA: Dual Temporal-channel-wise Attention for Spiking Neural Networks

Minje Kim[1*] Minjun Kim[2*] Xu Yang[2†]
[1]Promedius Inc.
[2]Beijing Institute of Technology
iankimrok@gmail.com, mnjnkai@gmail.com, pyro_yangxu@bit.edu.cn

Abstract

Spiking Neural Networks (SNNs) present a more energy-efficient alternative to Artificial Neural Networks (ANNs) by harnessing spatio-temporal dynamics and event-driven spikes. Effective utilization of temporal information is crucial for SNNs, leading to the exploration of attention mechanisms to enhance this capability. Conventional attention operations either apply identical operation or employ non-identical operations across target dimensions. We identify that these approaches provide distinct perspectives on temporal information. To leverage the strengths of both operations, we propose a novel Dual Temporal-channel-wise Attention (DTA) mechanism that integrates both identical/non-identical attention strategies. To the best of our knowledge, this is the first attempt to concentrate on both the correlation and dependency of temporal-channel using both identical and non-identical attention operations. Experimental results demonstrate that the DTA mechanism achieves state-of-the-art performance on both static datasets (CIFAR10, CIFAR100, ImageNet-1k) and dynamic dataset (CIFAR10-DVS), elevating spike representation and capturing complex temporal-channel relationship. We open-source our code: https://github.com/MnJnKIM/DTA-SNN.

1. Introduction

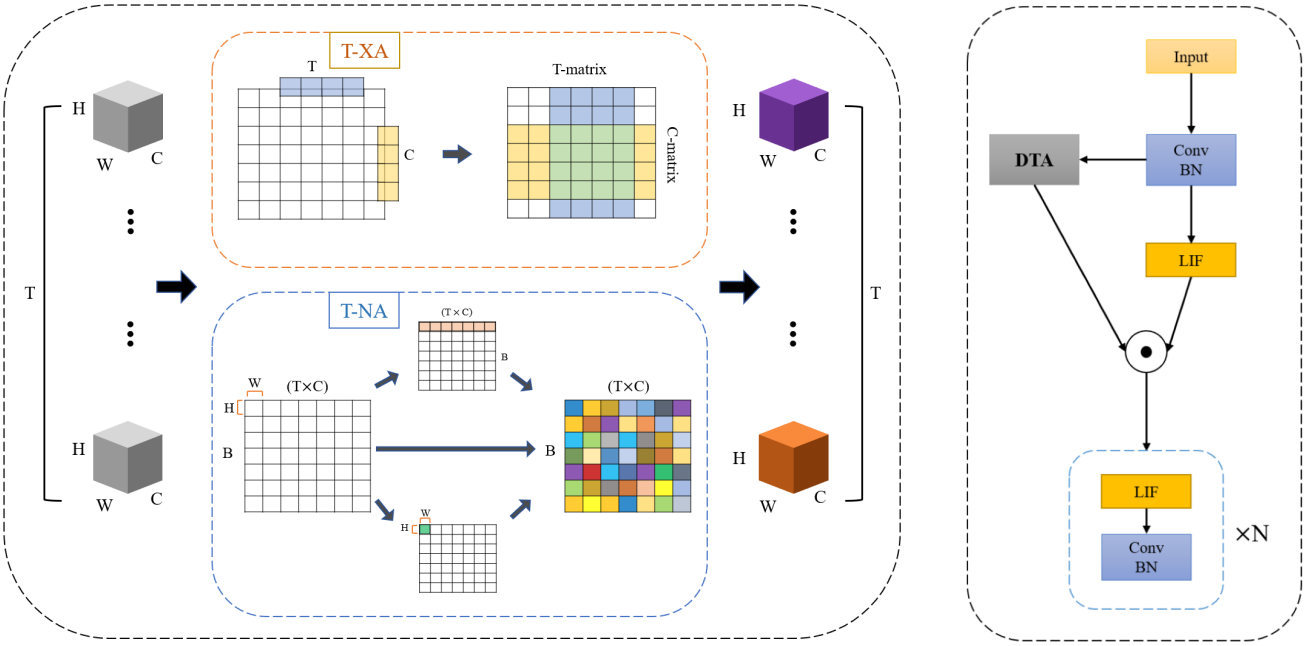
Artificial Neural Networks (ANNs) have achieved significant strides in image recognition through attention mechanisms and complex architectures like deeper-wider neural networks. However, these approaches accompany a substantial increase in computational demands and power consumption, which poses challenges for practical applications. To alleviate these challenges, Spiking Neural Networks (SNNs) have emerged, leveraging spatio-temporal dynamics and event-based spikes as activation functions. SNNs replace power-intensive multiply-accumulate operations with more efficient accumulation processes [37]. These characteristics enable SNNs to more accurately resemble the actual computations of the human brain [32], offering extreme energy efficiency and low-latency calculations when implemented in neuromorphic hardware [3, 12, 33].

Nevertheless, the spike-based information transmission poses significant challenges due to their non-differentiable activation function. To address this issue, the ANN-to-SNN (A2S) conversion method [40] establishes a connection between the neurons of SNNs and ANNs, enabling ANNs to be trained first and then converted to SNNs. While A2S conversion achieves comparable performance to ANNs, it requires numerous time steps and disregards the various alterations of temporal information in SNNs. In contrast, direct training approaches [2, 34, 44], which train SNNs using surrogate gradient (SG) alleviate the non-differentiable for spike activation. SG-based methods [4, 9, 29] demonstrate effective performance on large datasets with fewer timesteps and can directly process temporal data.

To further advance the capability of SNNs in selectively attending to pertinent information within temporal data, attention mechanism has seen widespread adoption. For instance, TA [46] demonstrates the potential of temporal-wise attention for SNNs. Multi-dimensional Attention methodology [48] enhances performance through the non-identical attention across temporal, channel, and spatial dimensions. Additionally, TCJA [55] effectively extracts spatio-temporal features by combining temporal and channel information with an identical attention at the same stage. Recently, Gated Attention Coding (GAC) [35] leverages non-identical attention with a single multi-dimensional attention block at the input stage.

In this paper, we consider that the attention mechanism depends on the identical and non-identical operation, which yield different expressive capabilities, and both operations improve spike representation. To realize this, we propose a novel dual Temporal-channel-wise attention (DTA) mechanism, which efficiently and effectively enhances feature representation for SNNs by combining the benefits of both

*Equal contribution.
†Corresponding author.



(a) DTA architecture. (b) DTA-SNN architecture.

Figure 1. (a) shows the overall workflow of our proposed Dual Temporal-channel-wise Attention (DTA) mechanism. The DTA incorporates Temporal-channel-wise identical Cross Attention (T-XA) and Temporal-channel-wise Non-identical Attention (T-NA) modules to administer an identical operation and non-identical operations across temporal and channel dimensions, respectively. (b) illustrates our proposed DTA-SNN architecture, which embeds a single DTA block into a spiking neural network (SNN).

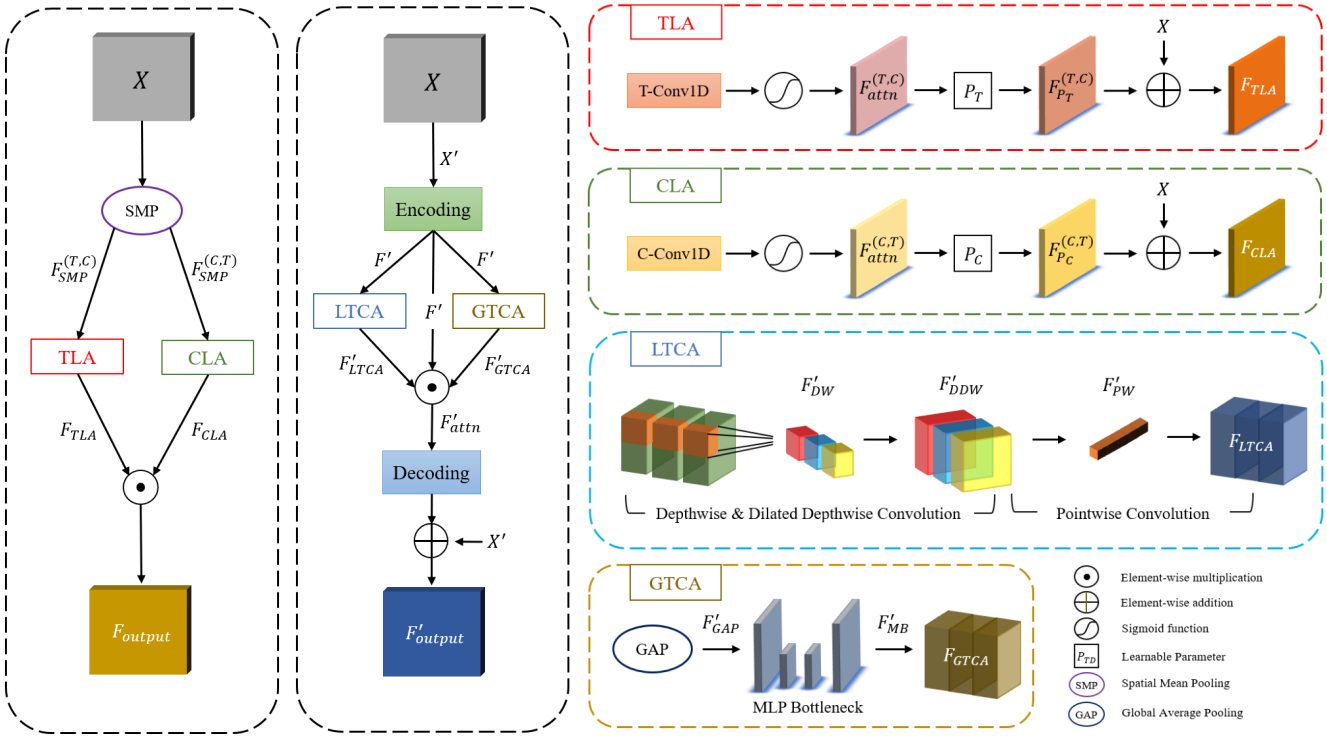
identical and non-identical attention operations, leading to seizure of complex temporal-channel dependency. As illustrated in Figure 1a, the DTA mechanism is composed of two independent attention strategies: Temporal-channel-wise identical Cross Attention (T-XA) and Temporal-channel-wise Non-identical Attention (T-NA). Initially, the T-XA module performs identical operation of temporal and channel dimensions to ensure elaborate temporal-channel correlation via cross attention. In contrast, the T-NA module addresses non-identical operations to interpret both local and global dependencies of the temporal-channel, leading to abundant spike representation. Additionally, as visualized in Figure 1b, we implement SNNs consisting of a single DTA block by adopting MS-ResNet structure [20] and GAC scheme [35] to prove that a single DTA block can outperform previous studies using multiple attention blocks in SNNs. Our implementation alleviates the high computational cost, memory usage, and low interpret-ability associated with multiple attention blocks. Our contributions are summarized below:

- To the best of our knowledge, this is the first attempt to incorporate both identical/non-identical attention mechanisms into the SNNs. Our proposed DTA mechanism notes that the expressive capabilities of attention relies on a dual temporal-channel-wise perspective.
- We introduce a novel T-XA module, which simultaneously considers the correlation of temporal and channel information with the identical attention operation. We also show that the T-NA module employs non-identical operations to handle the combined temporal-channel dimension from both intra- and inter-dependencies.
- Our proposed DTA mechanism, incorporating a single DTA block, achieves state-of-the-art performance on static and dynamic datasets. Experimental results demonstrate exceptional accuracy, with 96.73% /81.16% on CIFAR10/100 [1], 71.29% on ImageNet-1k [25], and 81.3% on CIFAR10-DVS [27].

2. Related Works

2.1. SNN Training Methods

The two primary training approaches for SNNs are the A2S conversion and the DT of SNNs. The A2S conversion methods typically involve transforming pre-trained ANNs into SNNs by replacing the activation functions with spiking neurons. For instance, these conversion methods have been proposed, including threshold balancing [6], spiking equivalents [38], soft reset [17], calibration algorithms [19, 28], rate norm layer [7], quantization [21, 26, 42], and



(a) T-XA module. (b) T-NA module. (c) Details of T-XA and T-NA modules.

Figure 2. (a) presents the overall structure of the T-XA module, (b) illustrates the design of the T-NA module, and (c) shows the overall architecture of the Temporal Local Attention (TLA), the Channel Local Attention (CLA), the Local Temporal-Channel Attention (LTCA), and the Global Temporal-Channel Attention (GTCA).

residual membrane potential [18, 42]. While these methods are predominantly applied to CNNs, recent works [24, 50] have explored their application to Transformers. However, A2S conversion methods have not yet surpassed the performance of ANNs, and the rate coding scheme of these methods constrains the temporal dynamics of SNNs, leading to elevated energy consumption.

The DT approaches [9, 34] train SNNs from scratch and utilize SG to resolve non-differentiability of spiking neurons. These methods are capable of handling temporal data, such as event-based datasets, with fewer time-steps required for training. Conventional spiking neuron models, including integrate-and-fire (IF) and leaky integrate-and-fire (LIF) [44], which implements spike-based information transmission, while recent advancements introduced novel models such as PLIF [10], GLIF [49], and CLIF [22]. Early DT methods, like BPTT [34], iteratively update gradients through spatial and temporal dimensions, and [14, 45] demonstrated efficiency on both dynamic and static datasets. However, these methods adopted SG as a tanh-like function, causing explosion or vanishing of gradient. Recent researches addressed these challenges by proposing various SG methods, including triangular shapes [4, 29, 36], sigmoid functions [43, 51], and adaptive-learnable SG [5, 30, 41]. Additionally, several studies [8, 23, 52] explored temporal-based batch normalization to manage temporal covariate shift.

2.2. Attention Mechanism in SNNs

The attention mechanism in CNNs overcomes inherent limitations by focusing on salient regions or channels within an image and selectively emphasizing prominent features. This advantage has catalyzed research into extending these techniques to SNNs, investigating how attention mechanisms can be harnessed to enhance performance of SNNs. TA [46] presented a temporal-wise attention mechanism with the squeeze-and-excitation block, which assigns attention factors to each temporal-wise input frame in SNNs, and demonstrated the potential of attention mechanisms in processing temporal data. The multi-dimensional attention methodology [48] promotes membrane potentials to use attention weights via sequential application of temporal/channel/spatial-wise attention, and their performance gains were validated across various benchmark datasets. TCJA [55] has proven the effectiveness of joint attention by integrating temporal and channel information with 1-D convolution, enabling the low-cost extraction of spatio-temporal features in SNNs. Nonetheless,

the aforementioned methods rely on direct coding, which necessitates the use of multiple attention blocks. Consequently, these methods frequently generate iterative-similar outputs at each time step, resulting in weak spike representation and limited spatio-temporal dynamics. To address this issue, Gated Attention Coding (GAC) [35] effectively captured variations in temporal information using a single attention block and demonstrated strong performance on static datasets.

3. Methods

3.1. Neuron Model and Surrogate Gradient for SNNs

We adopt the iterative LIF neuron model proposed in [44], and its dynamics can be described as follows:

$$u^n(t) = \tau u^n(t-1) \odot (1 - s^n(t-1)) + c^n(t), \quad (1)$$

where $u^n(t)$ represents the membrane potential of spiking neuron in the n-th layer at the time step t, τ is the time constant determining the decay rate of membrane potential, and $(1 - s^n(t-1))$ resets the potential to 0 after a spike. The symbol $\odot$ denotes element-wise multiplication.

The synaptic input current $c^n(t)$ is computed as:

$$c^n(t) = W^n \star s^{n-1}(t), \quad (2)$$

where W^n is the synaptic weight matrix, and $s^{n-1}(t)$ represents the output spikes from the previous layer at the time step t. The symbol $\star$ denotes the synaptic operation, which involves either convolutional or fully connected layers.

When the membrane potential $u^n(t)$ exceeds the threshold, a spiking output $s^n(t)$ is generated. It is defined as follows:

$$s^n(t) = \mathbb{H}(u^n(t) - V_{th}), \quad (3)$$

where $s^n(t)$ is the binary spiking output generated by the Heaviside step function $\mathbb{H}$, and V_{th} is the firing threshold. The Heaviside step function yields 1 when $(u^n(t) - V_{th})$ is non-negative, and 0 otherwise. Additionally, we update the weight of SNNs using the spatial-temporal backpropagation [44]:

$$\Delta W^n = \sum_t \frac{\partial L}{\partial W^n} = \sum_t \frac{\partial L}{s^n(t)} \frac{\partial s^n(t)}{\partial u^n(t)} \frac{\partial u^n(t)}{\partial c^n(t)} \frac{\partial c^n(t)}{\partial W^n}, \quad (4)$$

From the above equation, the term $\partial s^n(t)/\partial u^n(t)$ indicates gradient of the spiking function, which remains zero value with the exception of the scenario where $u^n(t)$ equals to Vth. To circumvent the non-differentiability, previous studies have introduced diverse forms of SG methods [4, 9, 29]. We implement the triangular SG method, as follows:

$$\frac{\partial s^n(t)}{\partial u^n(t)} = \begin{cases} \alpha(1 - \alpha \cdot \delta^n(t)) & \text{if} \quad \delta^n(t) < \frac{1}{\alpha} \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

where the constant α determines the maximum activation range of the gradient by adjusting the non-zero interval, and $\delta^n(t)$ denotes $|u^n(t) - V_{th}|$.

3.2. Dual Temporal-channel-wise Attention

We present the Dual Temporal-channel-wise Attention (DTA) mechanism, which integrates two identical/non-identical attention strategies: Temporal-channel-wise identical Cross Attention (T-XA) and Temporal-channel-wise Non-identical Attention (T-NA). Our proposed DTA mechanism aims to efficiently and richly enhance feature representation by combining the benefits of both identical and non-identical attention, effectively capturing complex temporal-channel correlation and temporal-channel dependency:

$$O_{DTA} = \sigma(O_{T-XA} \odot O_{T-NA}) \odot Spikes, \quad (6)$$

The outputs of the DTA block are derived by combining the outputs of T-XA and T-NA through element-wise multiplication $\odot$ with the sigmoid function σ. This process leverages the strengths of both identical and non-identical attention mechanisms, by simultaneously integrating the emphasis with spikes. Specifically, as shown in Figure 1a, the T-XA module executes an identical operation across both temporal and channel dimensions, ensuring fine-grained temporal-channel correlation. Contrastively, the T-NA module addresses inputs by applying non-identical operations that comprehend both local and global dependencies across the temporal and channel dimensions, enabling abundant feature representation in SNNs.

3.3. Temporal-channel-wise identical Cross Attention

We first use the T-XA module to refine XA in the DTA block, as shown in Figure 2a. The T-XA module is a global attention branch, which consists of Temporal-wise Local Attention ($TLA(\cdot)$) and Channel-wise Local Attention ($CLA(\cdot)$) branches. To implement the T-XA module, we consider the correlation of temporal and channel information through cross attention and the operation of the T-XA module, as follows:

$$T{-}XA(X) = TLA(X) \odot CLA(X), \quad (7)$$

where $X \in \mathbb{R}^{(T \times C \times H \times W)}$ is the inputs from pre-synaptic neurons. We transform the X into enhanced features $F_{SMP}^{(TD,SD)} \in \mathbb{R}^{(TD \times SD)}$. Here, TD and SD represent that the target and subtarget dimensions, respectively, and we primarily focus on TD, considering SD in the each local branch. Additionally, we use Spatial Mean Pooling (SMP) to capture the global cross-acceptance field, which reflects interactive correlation, supplying efficiently extracted features. The $F_{SMP}^{(TD,SD)}$ is generated through SMP over the

spatial dimensions H and W of the X, and the $F_{SMP}^{(TD,SD)}$ is defined as:

$$F_{SMP}^{(TD,SD)} = SMP(X, TD, SD), \\ = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X^{(TD,SD)}. \quad (8)$$

Each local attention branch simultaneously processes the temporal and channel dimensions while using a small number of optimized parameters to provide attention, thereby enriching the temporal-channel representation in SNNs. As shown in Figure 2c, We obtain the features F_{LA} from the local attention branch and the definition is as follows:

$$F_{LA} = P_{TD}(\sigma(conv(F_{SMP}^{(TD,SD)}))) \oplus X^{(TD,SD)}, \quad (9)$$

where the $conv(\cdot)$ is 1D convolution operation along the TD with the SD, the feature map F^{TD} obtained by this operation is generated through the sigmoid function σ, and the attention map F_{attn}^{TD} is scaled with the learnable parameter P_{TD}. Finally, F_{attn}^{TD} and $F^{(TD,SD)}$ are combined through the residual operation $\oplus$ to obtain F_{LA}. F_{TLA} and F_{CLA}, obtained from $TLA(\cdot)$ and $CLA(\cdot)$ respectively, are then merged using the element-wise multiplication operation to output features F_{output} of T-XA.

3.4. Temporal-channel-wise Non-identical Attention

As illustrated in Figure 2b, we introduce the T-NA module to effectively address both intra/inter-dependencies between temporal and channel information via non-identical operations. Concretely, our T-NA module consists of local and global temporal-channel attention, and given inputs $X \in \mathbb{R}^{T\times C\times H\times W}$ are reshaped inputs $X' \in \mathbb{R}^{(T\times C)\times H\times W}$ for alleviating computational cost of attention such as 3D convolution or 3D pooling. The X' is projected to attain features F' via encoding operation for the enriched representation of X':

$$F' = f(X'), \quad (10)$$

where functional operation $f(\cdot)$ includes 2D convolution operation with 1×1 kernel and GELU function.

As shown in Figure 2c, to address intra-dependency of temporal-channel information, we propose the local temporal-channel attention (LTCA) mechanism, which processes the F' using several convolution operations as follows:

$$LTCA(F') = f_{PW}(f_{DDW}(f_{DW}(X'))), \quad (11)$$

where f_{PW}, f_{DDW} and f_{DW} are point-wise, dilation-depth-wise and depth-wise convolution operations, respectively, and those convolution operations effectively capture intra-dependency of the temporal-channel while maintaining computational efficiency. Concurrently, we present the global temporal-channel attention (GTCA) mechanism to enable adaptive responses to inter-information of the temporal-channel, as follows:

$$GTCA(F') = f_{MB}(f_{GAP}(F')), \quad (12)$$

where $f_{GAP}(\cdot)$ indicates global average pooling, which compresses F' by averaging spatial dimensions to highlight the global context of the temporal-channel features. Then, the emphasized features are manipulated using the MLP bottleneck structure $f_{MB}(\cdot)$, which consists of the $Linear{-}ReLU{-}Linear$ sequence. This f_{MB} first squeezes and then expands them in order to reinforce the feature representation by effectively incorporating inter-dependency of temporal-channel information. Subsequently, the attention maps from $LTCA(\cdot)$ and $GTCA(\cdot)$ are integrated through element-wise multiplication with the original F' to yield the attention features F'_{attn}:

$$F'_{attn} = F'_{LTCA} \odot F'_{GTCA} \odot F'. \quad (13)$$

Finally, we exploit decoding operation $f_{conv}(\cdot)$ with 1×1 convolution to handle the F'_{attn}, and utilize residual connection for facilitating more accurate attention and contributing more stable training in the T-NA module:

$$F'_{output} = f_{conv}(F'_{attn}) \oplus X'. \quad (14)$$

Layer	C_{Out}	I_{Out}	ResNet-18	ResNet-34
Conv1	32×32	112×112	3×3, 64, s=1	7×7, 64, s=2
Conv2	32×32	56×56	[3×3, 128] [3×3, 128] ×3	[3×3, 64] [3×3, 64] ×3
Conv3	16×16	28×28	[3×3, 256] [3×3, 256] ×3	[3×3, 128] [3×3, 128] ×4
Conv4	8×8	14×14	[3×3, 512] [3×3, 512] ×2	[3×3, 256] [3×3, 256] ×6
Conv5	-	7×7	-	[3×3, 512] [3×3, 512] ×3
	AveragePooling,		FC-10/100	FC-1000

Table 1. Architecture of MS-ResNet utilized for CIFAR10/100 and ImageNet-1k datasets in DTA-SNNs. C_{Out} and I_{Out} represent the sizes of output features in the CIFAR10/100 and ImageNet-1k datasets, respectively.

4. Experiments

In Section 4.1, we first describe the experimental setup used to evaluate our proposed DTA mechanism for SNNs. Next, in Section 4.2, we demonstrate the effectiveness of the

DTA mechanism through a comparative analysis with state-of-the-art (SOTA) methods. Finally, we conduct an ablation study to further investigate the impact of the individual modules of the DTA mechanism, as detailed in Section 4.3.

4.1. Settings

Datasets. We evaluate our DTA mechanism on four types of classification benchmark datasets, encompassing both static and dynamic datasets. First, we use CIFAR10 and CIFAR100 [1], widely recognized static image classification benchmarks with 10 and 100 classes of natural images, respectively. Next, we employ ImageNet-1k [25], a large-scale static image classification dataset comprising 1,000 classes of diverse objects. Finally, we utilize CIFAR10-DVS [27], an event-based image classification dataset derived from scanning each sample of the static CIFAR10 dataset using DVS cameras.

Dataset	Batch Size	Epochs	Time Step	Initial LR	Decay
CIFAR10	64	250	4/6	0.1	5e-5
CIFAR100	64	250	4/6	0.1	5e-5
ImageNet-1k	128	200	4/6	0.1	1e-5
CIFAR10-DVS	64	1000	10	0.05	0

Table 2. Hyper-parameter training settings for DTA-SNNs.

Implementation Details. We implement our proposed DTA mechanism using the PyTorch framework, conduct experiments on several NVIDIA A100 GPUs, and adopt the MS-ResNet structure [20] to build SNNs that consist of a single DTA block, as visualized in Figure 1b. For experiments on the CIFAR10/100, we utilize MS-ResNet18, whereas MS-ResNet34 is employed for the ImageNet-1k dataset, as shown in Table 1. In the case of the CIFAR10-DVS dataset, the input size is set to 48×48, which leads to adjustments in the internal output sizes of the MS-ResNet18 architecture. Furthermore, we use the SGD optimizer with 0.9 momentum and a cosine annealing schedule [31] for all our experiments, with detailed hyper-parameters described in Table 2. Additionally, the hyper-parameters for the iterative LIF neuron are configured with the firing threshold V_{th} of 1.0 and the time constant τ of 0.5. For the overall experiments, we follow widespread data augmentation strategies from previous studies [8, 16, 22, 23, 35, 47, 48, 53, 54]. Notably, in contrast to other SOTA algorithms [11, 22, 55] that use TET loss [4], we demonstrate the effectiveness of our attention mechanism using standard cross-entropy loss.

4.2. Comparisons with SOTA methods

CIFAR10/100. On the CIFAR10/100 datasets, each experiment was conducted three times, with the mean accuracy and standard deviation reported and summarized in Table 3. Our proposed method outperformed previous best results with higher accuracy and fewer time steps. Specifically, DTA-SNN achieves SOTA performance with 96.73% top-1 accuracy at 6 time steps and 96.50% accuracy at 4 time steps. Compared to non-identical attention method, DTA-SNN surpassed the GAC-SNN accuracy of 96.46% at 6 time steps with an accuracy of 96.50% at just 4 time steps on the CIFAR10 dataset. Moreover, we also obtained 0.71% higher performance than GAC-SNN at 6 time steps on the CIFAR100 dataset. These results demonstrate the effectiveness of our dual attention operations.

Method	Type	NN Architecture	Parameter(M)	Time Step	CIFAR10 Acc(%)	CIFAR100 Acc(%)
RMP-SNN [17]	A2S	VGG-16	33.64/34.01	2048	93.63	70.93
SRP [18]	A2S	VGG-16	33.64/34.01	32/64	95.42/95.40	77.01/77.10
QCFS [7]	A2S	VGG-16/ResNet-18	33.64/11.22	32/64	95.54/95.55	76.45/76.37
Diet-SNN [36]	DT	ResNet-20	-	10	92.54	64.07
Dspike [29]	DT	ResNet-18	11.17/11.22	6	94.25	74.24
STBP-tdBN [52]	DT	ResNet-19	12.63	4/6	92.92/93.16	-
TET [4]	DT	ResNet-19	12.63/12.67	4/6	94.44/94.50	74.47/74.72
TEBN [8]	DT	ResNet-19	12.63/12.67	4/6	95.58/95.60	76.13/76.41
GLIF [49]	DT	ResNet-19	12.63/12.67	4/6	94.85/95.03	77.05/77.35
Real Spike [15]	DT	ResNet-19/VGG-16	12.63/-	4/6	95.60/95.71	70.62/71.17
Ternary Spike [13]	DT	ResNet-19	12.63	1/2	95.28/95.60	78.13/79.66
TAB [23]	DT	ResNet-19	12.63/12.67	4/6	95.94/96.09	76.81/76.82
CLIF [22]	DT	ResNet-18	11.17/11.22	4/6	96.01/96.45	79.69/80.58
MPBN [16]	DT	ResNet-19	12.63/12.67	1/2	96.06/96.47	78.71/79.51
GAC-SNN [35]	DT	MS-ResNet-18	12.63/12.67	4/6	96.24/96.46	79.83/80.45
DTA-SNN(Ours)	DT	MS-ResNet-18	12.99/13.03	4	**96.50 ± 0.09**	**79.94 ± 0.08**
				6	**96.73 ± 0.11**	**81.16 ± 0.27**

Table 3. Comparison with SOTA on CIFAR10/100.

Method	Type	NN Architecture	Parameter(M)	Time Step	Accuracy(%)
SRP [18]	A2S	Modified-VGG-16	138.36	32/64	69.35/69.43
QCFS [7]	A2S	ResNet-34	21.79	32	69.37
RMP-SNN [17]	A2S	ResNet-34	21.79	4096	69.89
Diet-SNN [36]	DT	VGG-16	-	5	69.00
Dspike [29]	DT	ResNet-34	21.79	6	68.19
TET [4]	DT	ResNet-34	21.79	6	64.79
SEW-ResNet [9]	DT	SEW-ResNet-34	21.79	4	67.04
MS-ResNet [20]	DT	Ms-ResNet-34	21.80	6	69.43
GLIF [49]	DT	ResNet-34	21.79	6	69.09
Real Spike [15]	DT	ResNet-34	21.79	4	67.69
Ternary Spike [13]	DT	ResNet-34	21.79	4	70.12
TAB [23]	DT	ResNet-34	21.79	4	67.78
MPBN [16]	DT	ResNet-34	21.79	4	64.71
GAC-SNN [35]	DT	MS-ResNet-34	21.93	4/6	69.77/70.42
DTA-SNN(Ours)	DT	MS-ResNet-34	22.02	4 6	**70.27** **71.29**

Table 4. Comparison with SOTA methods on ImageNet-1k.

Method	NN Architecture	T	Accuracy (%)
Dspike [29]	ResNet-18	10	75.4
STBP-tdBN [52]	ResNet-19	4	67.8
TET [4]	VGGSNN	10	77.3
TEBN [8]	7-layerCNN	10	75.1
SEW-ResNet [9]	SEW-ResNet	16	74.4
MS-ResNet [20]	Ms-ResNet-20	-	75.6
GLIF [49]	7B-wideNet	16	78.1
Real Spike [15]	ResNet-20	10	78.0
Ternary Spike [13]	ResNet-20	10	78.7
TAB [23]	7-layerCNN	4	76.7
MPBN [16]	ResNet-20	10	78.7
Spikformer [54]	Spikformer-2-256	10	78.9
Spikingformer [53]	Spikingformer-2-256	10	79.9
Spike-driven [47]	S-Transformer-2-256	16	80.0
TA-SNN [46]	5-layerCNN	10	72.0
TCJA-SNN [55]	VGGSNN	10	80.7
DTA-SNN(Ours)	MS-ResNet-18	10	**81.3± 0.3**

Table 5. Comparison with SOTA methods on CIFAR10-DVS.

ImageNet-1k. We evaluated our DTA mechanism using the widely utilized large-scale static dataset ImageNet-1k and adopted MS-ResNet34 as the backbone, assessing the mechanism at 4 and 6 time steps, as detailed in Table 4. At 4 time steps, we achieved a notable performance of 70.27%. Our method, regardless of the training type, outperformed recent A2C conversion methods with substantially fewer time steps: SRP (69.43%), QCFS (69.37%), RMP-SNN (69.89%), and DT methods: Dspike (68.19%) and GLIF (69.09%). At 6 time steps, we achieved an additional 0.87% increase in top-1 accuracy, surpassing the best result of previous attention mechanisms [35] at the same time step, with only a minor parameter increase of approximately 0.09. Experiments on large-scale static datasets demonstrate that the method consistently achieved strong performance regardless of the training type, approach, or number of time steps.

CIFAR10-DVS. We evaluate the proposed DTA mechanism on the widely used dynamic dataset CIFAR10-DVS, which provides 0.9k training samples per label. Compared to static datasets, dynamic datasets suffer from significant noise, making them more prone to overfitting when trained with complex architectures. However, our approach, utilizing MS-ResNet-18 as the backbone, achieves best with an accuracy of 81.03% at 10 time steps, surpassing previous approaches based on transformer architectures [47, 53, 54]. This demonstrates that our model effectively learns temporal patterns in dynamic datasets. The results of our experiments on CIFAR10-DVS are reported as the mean and standard deviation over three runs, as shown in Table 5.

4.3. Ablation Study

We conducted experiments on both the static CIFAR100 dataset and the dynamic CIFAR10-DVS dataset to validate the effectiveness of our proposed identical and non-identical attention approach, built upon temporal-channel-wise processing. To comprehensively assess the impact of the T-XA/T-NA modules and the DTA block, we performed a series of ablation studies. The results, detailed in Table 6, demonstrate that the DTA mechanism is crucial for boosting overall performance. The DTA mechanism achieved the highest performance, with improvements of approximately 1.6% and 1.2% on CIFAR100 and CIFAR10-DVS, respectively, compared to the baseline. This indicates that applying the attention mechanism to enrich spike representation in SNNs is reasonable. Furthermore, the inherent dynamics in SNNs, driven by temporal information, resulted in slightly more noticeable performance gains from attention

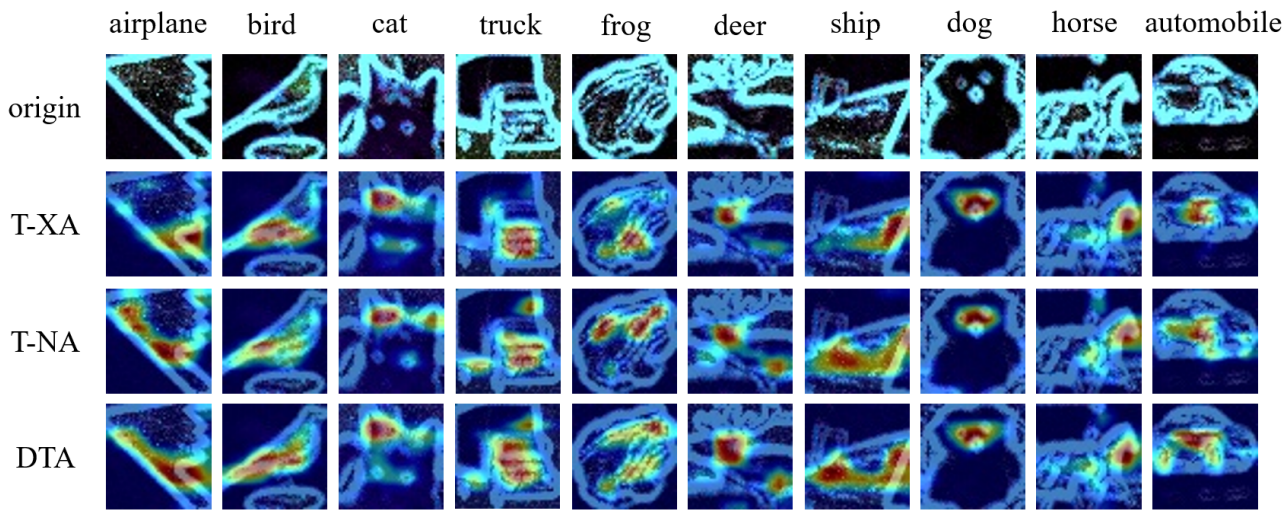


Figure 3. Grad-CAM results for CIFAR10-DVS, visualized across 10 classes in four rows. The first row shows the original images, followed by Grad-CAM visualizations generated using only T-XA in the second row and only T-NA in the third row. The fourth row illustrates the results of the DTA mechanism.

T-XA	T-NA	CIFAR100 (T=6)	CIFAR10-DVS (T=10)
✗	✗	79.74	80.4
✓	✗	80.56	80.8
✗	✓	80.78	81.2
✓	✓	**81.28**	**81.6**

Table 6. Top-1 test accuracy (%) for ablation studies of T-XA and T-NA modules.

mechanisms in the dynamic dataset (CIFAR10-DVS) compared to the static dataset (CIFAR100), as observed across all ablation studies. This implies that the attention mechanism, which mirrors the dynamic neural processes in humans, exerts a more pronounced effect in dynamic environments. Moreover, in most SNN architectures, the number of simulation time steps is generally higher than the number of channels, leading us to interpret this as a need to emphasize temporal-channel relationship through attention. Therefore, we considered both identical and non-identical attention mechanisms for temporal-channel correlation and dependency. As shown in the experimental results, the application of each module individually outperformed the baseline performance. Thus, the DTA combined from each module yielded the most optimal results, capturing a broader range of relevant features compared to any module used in isolation. In addtion, as shown in Figure 3, we deployed Grad-CAM [39] on 10 classes to assess the impact of removing specific components within the DTA-SNN architecture. The study highlights how our proposed attention mechanism affects the model's ability to localize key features even in dynamic data with significant noise. In summary, this demonstrates the remarkable effectiveness of the complementary dual attention operations, which accounts for both correlation and dependency within the temporal-channel domain.

5. Conclusion

In this work, we introduce a novel DTA mechanism that integrates the T-XA and the T-NA modules. Our proposed DTA mechanism addresses a gap in prior research by simultaneously exploiting both identical and non-identical attention operations to analyze temporal-channel correlation and dependency. In detail, the T-XA module focuses on temporal-channel correlations through an identical attention operation, while the T-NA module captures both local and global dependencies of the temporal-channel using non-identical attention operations. By consolidating these two strategies into a single DTA block, our proposed attention mechanism effectively enriches the spike representation and identifies variations in temporal-channel information, enhancing the comprehension of SNNs regarding complex temporal-channel relationship. Thus, extensive experiments demonstrate that our method consistently yields SOTA performance on both static and dynamic datasets, including CIFAR10 (96.73%), CIFAR100 (81.16%), ImageNet-1k (71.29%), and CIFAR10-DVS (81.3%).

Acknowledgement

This research was funded by the STI 2030-Major Projects 2022ZD0208700, and the Fundamental Research Funds for the Central Universities.

References

[1] A.Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, Univ. Toronto, 2009. 2, 6

[2] Sander M Bohte, Joost N Kok, and Johannes A La Poutré. Spikeprop: backpropagation for networks of spiking neurons. In *ESANN*, volume 48, pages 419–424. Bruges, 2000. 1

[3] 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

[4] Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. In *International Conference on Learning Representations*, 2022. 1, 3, 4, 6, 7

[5] Shikuang Deng, Hao Lin, Yuhang Li, and Shi Gu. Surrogate module learning: Reduce the gradient error accumulation in training spiking neural networks. In *International Conference on Machine Learning*, pages 7645–7657. PMLR, 2023. 3

[6] Peter U Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *2015 International joint conference on neural networks (IJCNN)*, pages 1–8. ieee, 2015. 2

[7] Jianhao Ding, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Optimal ann-snn conversion for fast and accurate inference in deep spiking neural networks. *arXiv preprint arXiv:2105.11654*, 2021. 2, 6, 7

[8] Chaoteng Duan, Jianhao Ding, Shiyan Chen, Zhaofei Yu, and Tiejun Huang. Temporal effective batch normalization in spiking neural networks. *Advances in Neural Information Processing Systems*, 35:34377–34390, 2022. 3, 6, 7

[9] 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, 3, 4, 7

[10] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothee 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 (ICCV)*, pages 2661–2671, October 2021. 3

[11] Wei Fang, Zhaofei Yu, Zhaokun Zhou, Ding Chen, Yanqi Chen, Zhengyu Ma, Timothée Masquelier, and Yonghong Tian. Parallel spiking neurons with high efficiency and ability to learn long-term dependencies. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[12] Steve Furber. Large-scale neuromorphic computing systems. *Journal of neural engineering*, 13(5):051001, 2016. 1

[13] Yufei Guo, Yuanpei Chen, Xiaode Liu, Weihang Peng, Yuhan Zhang, Xuhui Huang, and Zhe Ma. Ternary spike: Learning ternary spikes for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12244–12252, 2024. 6, 7

[14] 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. 3

[15] 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, 7

[16] 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. 6, 7

[17] 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, 6, 7

[18] Zecheng Hao, Tong Bu, Jianhao Ding, Tiejun Huang, and Zhaofei Yu. Reducing ann-snn conversion error through residual membrane potential. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11–21, 2023. 3, 6, 7

[19] Zecheng Hao, Jianhao Ding, Tong Bu, Tiejun Huang, and Zhaofei Yu. Bridging the gap between anns and snns by calibrating offset spikes. *arXiv preprint arXiv:2302.10685*, 2023. 2

[20] Yifan Hu, Lei Deng, Yujie Wu, Man Yao, and Guoqi Li. Advancing spiking neural networks toward deep residual learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2, 6, 7

[21] Yangfan Hu, Qian Zheng, Xudong Jiang, and Gang Pan. Fast-snn: fast spiking neural network by converting quantized ann. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[22] Yulong Huang, Xiaopeng LIN, Hongwei Ren, Haotian FU, Yue Zhou, Zunchang LIU, biao pan, and Bojun Cheng. CLIF: Complementary leaky integrate-and-fire neuron for spiking neural networks. In *Forty-first International Conference on Machine Learning*, 2024. 3, 6

[23] Haiyan Jiang, Vincent Zoonekynd, Giulia De Masi, Bin Gu, and Huan Xiong. TAB: Temporal accumulated batch normalization in spiking neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 6, 7

[24] Yizhou Jiang, Kunlin Hu, Tianren Zhang, Haichuan Gao, Yuqian Liu, Ying Fang, and Feng Chen. Spatio-temporal approximation: A training-free SNN conversion for transformers. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 2, 6

[26] Chen Li, Lei Ma, and Steve Furber. Quantization framework for fast spiking neural networks. *Frontiers in Neuroscience*, 16:918793, 2022. 2

[27] 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. 2, 6

[28] Yuhang Li, Shikuang Deng, Xin Dong, Ruihao Gong, and Shi Gu. A free lunch from ann: Towards efficient, accurate spiking neural networks calibration. In *International conference on machine learning*, pages 6316–6325. PMLR, 2021. 2

[29] Yuhang Li, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, and Shi Gu. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. *Advances in Neural Information Processing Systems*, 34:23426–23439, 2021. 1, 3, 4, 6, 7

[30] Shuang Lian, Jiangrong Shen, Qianhui Liu, Ziming Wang, Rui Yan, and Huajin Tang. Learnable surrogate gradient for direct training spiking neural networks. In *IJCAI*, pages 3002–3010, 2023. 3

[31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[32] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997. 1

[33] Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673, 2014. 1

[34] 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. 1, 3

[35] Xuerui Qiu, Rui-Jie Zhu, Yuhong Chou, Zhaorui Wang, Liang-jian Deng, and Guoqi Li. Gated attention coding for training high-performance and efficient spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 601–610, 2024. 1, 2, 4, 6, 7

[36] Nitin Rathi and Kaushik Roy. Diet-snn: Direct input encoding with leakage and threshold optimization in deep spiking neural networks. *arXiv preprint arXiv:2008.03658*, 2020. 3, 6, 7

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

[38] Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017. 2

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[40] Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95, 2019. 1

[41] Ziming Wang, Runhao Jiang, Shuang Lian, Rui Yan, and Huajin Tang. Adaptive smoothing gradient learning for spiking neural networks. In *International Conference on Machine Learning*, pages 35798–35816. PMLR, 2023. 3

[42] Ziming Wang, Yuhao Zhang, Shuang Lian, Xiaoxin Cui, Rui Yan, and Huajin Tang. Toward high-accuracy and low-latency spiking neural networks with two-stage optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2, 3

[43] Stanisław Woźniak, Angeliki Pantazi, Thomas Bohnstingl, and Evangelos Eleftheriou. Deep learning incorporating biologically inspired neural dynamics and in-memory computing. *Nature Machine Intelligence*, 2(6):325–336, 2020. 3

[44] 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, 4

[45] 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*, volume 33, pages 1311–1318, 2019. 3

[46] 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. 1, 3, 7

[47] Man Yao, JiaKui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo XU, and Guoqi Li. Spike-driven transformer. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 6, 7

[48] 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. 1, 3, 6

[49] Xingting Yao, Fanrong Li, Zitao Mo, and Jian Cheng. GLIF: A unified gated leaky integrate-and-fire neuron for spiking neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 3, 6, 7

[50] Kang You, Zekai Xu, Chen Nie, Zhijie Deng, Xiang Wang, Qinghai Guo, and Zhezhi He. Spikezip-tf: Conversion is all you need for transformer-based snn. In *Forty-first International Conference on Machine Learning (ICML)*, 2024. 3

[51] Friedemann Zenke and Tim P Vogels. The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *Neural computation*, 33(4):899–925, 2021. 3

[52] 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*, volume 35, pages 11062–11070, 2021. 3, 6, 7

[53] 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, 2023. 6, 7

[54] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng YAN, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representations*, 2023. 6, 7

[55] 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. 1, 3, 6, 7