MSVIT: Improving Spiking Vision Transformer Using Multi-scale Attention Fusion

Wei Hua[1], **Chenlin Zhou**[2], **Jibin Wu**[3], **Yansong Chua**[1*] and **Yangyang Shu**[4*]
[1]China Nanhu Academy of Electronics and Information Technology, China
[2]University of Chinese Academy of Sciences, China
[3]The Hong Kong Polytechnic University, Hong Kong SAR, China
[4]School of Systems and Computing, The University of New South Wales, Australia
{huawei, caiyansong}@cnaeit.com, zhouchenlin19@mails.ucas.ac.cn, jibin.wu@polyu.edu.hk,
yangyang.shu@unsw.edu.au

Abstract

The combination of Spiking Neural Networks (SNNs) with Vision Transformer architectures has garnered significant attention due to their potential for energy-efficient and high-performance computing paradigms. However, a substantial performance gap still exists between SNN-based and ANN-based transformer architectures. While existing methods propose spiking self-attention mechanisms that are successfully combined with SNNs, the overall architectures proposed by these methods suffer from a bottleneck in effectively extracting features from different image scales. In this paper, we address this issue and propose MSVIT. This novel spike-driven Transformer architecture firstly uses multi-scale spiking attention (MSSA) to enhance the capabilities of spiking attention blocks. We validate our approach across various main data sets. The experimental results show that MSVIT outperforms existing SNN-based models, positioning itself as a state-of-the-art solution among SNN-transformer architectures. The codes are available at https://github.com/Nanhu-AI-Lab/MSViT.

1 Introduction

Spiking Neural Networks (SNNs), referred to as third-generation neural networks [Maass, 1997], have garnered significant attention attributed to their biological plausibility, event-driven processing, and potential for high energy efficiency [Roy *et al.*, 2019; Pei *et al.*, 2019]. Despite these advantages, SNNs have yet to achieve performance levels comparable to traditional Artificial Neural Networks (ANNs), particularly in complex vision tasks. This performance gap presents a major obstacle to the widespread adoption of SNNs in practical applications.

Initially developed for natural language processing tasks, Transformers [Vaswani *et al.*, 2017] have been extensively explored and extended to various computer vision applications, including image classification [Dosovitskiy, 2020], object detection [Liu *et al.*, 2021], and semantic segmentation [Wang *et al.*, 2021]. Notably, the self-attention mechanism—the core component of Transformers enables the models to focus on salient input information selectively, bears a striking resemblance to the selective attention processes discovered in the human biological system [Whittington *et al.*, 2018].

Given the biological plausibility of self-attention mechanisms and their alignment with human cognitive processes, it is intuitive to investigate their integration within Spiking Neural Networks (SNNs) to enhance deep learning models. The combination of the powerful representational capabilities of Transformers with the energy-efficient nature of SNNs presents an exciting research direction. In recent years, several studies have explored incorporating spiking neurons into Transformer-based architectures. For instance, Zhou et al. [Zhou *et al.*, 2023] introduced "Spikformer," a spike-driven self-attention mechanism, marking the first attempt to integrate spike-driven neurons into the Transformer framework. Shi et al. [Shi *et al.*, 2024] proposed SpikingResformer, which combines ResNet-inspired architecture with self-attention computation to achieve competitive performance in image classification tasks while maintaining low energy consumption. Yao et al. [Yao *et al.*, 2024b; Yao *et al.*, 2024a] developed two versions of Spike-driven Transformer trained using sparse AND-ACcumulate (AC) operations, achieving state-of-the-art (SOTA) results on the ImageNet-1K dataset among SNN-based architectures in the same terms.

Despite these advancements, there remains a performance gap between SNN-based models and traditional Artificial Neural Network (ANN) counterparts. While spiking neurons offer energy-efficient processing, their binary nature (using only 0 and 1 spikes) poses significant challenges in training larger and deeper networks. Unlike ANNs, which rely on floating-point matrix multiplication and softmax operations, the binary representation in SNNs struggles to capture the intrinsic and diverse information present in input data, often resulting in suboptimal accuracy for downstream tasks. Addressing these limitations is critical to exploring the full potential of SNN-based Transformer architectures.

Multi-scale structures are extensively utilized across com-

*Corresponding authors.

puter vision (CV) [Fan *et al.*, 2021], natural language processing (NLP) [Guo *et al.*, 2020], and signal processing domains [Park *et al.*, 2019] due to their effectiveness in capturing patterns at varying scales. Studies on the visual cortex of cats and monkeys suggest that increasing the number of distinct channels, with each channel corresponding to progressively specialized features, enhances the model's ability to extract structured information from input images [Fan *et al.*, 2021; Koenderink, 1984]. These findings demonstrate that multi-scale feature extraction improves the robustness and generalizability of learned representations in CV tasks.

Motivated by these insights, we propose a novel spiking attention mechanism, MSSA, which incorporates Multi-Scale Spike Attention (MSSA) blocks into transformer architecture. Each attention head within the MSSA block operates at varying scales via multiple inputs, thereby enriching the perceptual field of the spiking self-attention mechanism. Larger scales capture more global and smoother features, while smaller scales focus on local details, enhancing the sharpness and distinctiveness of feature representations. The balance between global and local information of inputs is crucial for spiking neural networks (SNNs), where it is essential to develop innovative methods that expand the representational capacity of spike-based models.

Building on MSSA, we further introduce the Multi-Scale Spike-driven Transformer, comprising hierarchical layers with MSSAs. Each layer is designed with attention heads that process input images at different scales, enabling the network to capture multi-scale features effectively. This architecture bridges the seminal concept of multi-scale feature hierarchies with spike-driven Transformer models, leveraging principles of resolution and channel scaling to improve performance in various visual recognition tasks. We hypothesize that integrating multi-scale attention mechanisms into spike-driven Transformers will significantly enhance their capability to extract diverse and robust features, thereby advancing the state of spiking models in computer vision applications.

The contributions of this work are summarized as follows:

1. We develop a novel multi-scale spiking attention module, tailor-made for the SNN's attributes, which enables the spiking transformer to extract features from different scales of inputs by spikes with low energy costs, improving the performance of the spiking transformer.
2. We develop a direct-training hierarchical spiking transformer, namely MSVIT, incorporating MSSA into a vision transformer. This design marks the effective exploration of multi-scale spiking representation in Transformer-based SNNs.
3. We conduct extensive experiments on mainstream static and neuromorphic datasets, achieving state-of-the-art performance compared to the latest SNN-based models. Notably, MSVIT has surpassed QKFormer, which has achieved 84.22% top-1 accuracy on ImageNet-1K with 224^2 input size and 4-time steps using the direct training from scratch by 85.06%, positioning itself as a state-of-the-art solution among SNN-transformer architectures.

Finally, based on the aforementioned experimental results, we conduct the ablation study to discuss and analyze MSVIT. The source code is open-sourced and available at https://github.com/Nanhu-AI-Lab/MSViT.

2 Related Work

2.1 Vision Transformers

ViTs segment images into patches and apply self-attention [Vaswani *et al.*, 2017; Devlin, 2018] to learn contextual relationships, effectively reducing inductive bias [Neil and Dirk, 2020; Neil and Dirk, 2020] and outperforming CNNs across multiple vision tasks [Mei *et al.*, 2021; Bertasius *et al.*, 2021; Guo *et al.*, 2021]. Nevertheless, ViTs face challenges like high parameter counts [Guo *et al.*, 2021], and increased computational complexity proportional to token length [Pan *et al.*, 2020; Liu *et al.*, 2022]. To enhance the computational efficiency of ViTs, many researchers [Jie and Deng, 2023; Li *et al.*, 2023] focused on exploring lightweight improvement methods from transformer architectures. For example, LeViT [Graham *et al.*, 2021] incorporates convolutional elements to expedite processing, and MobileViT [Mehta and Rastegari, 2021] combines lightweight MobileNet blocks with MHSA, achieving lightweight ViTs successfully. However, these enhancements still rely on expensive MAC computations, which are not suitable for Edge devices. This highlights the need for investigating more energy-efficient ViT solutions. Involving SNNs in Transformer architectures is one of the approaches.

2.2 Transformer Architecture in Spiking Neural Networks

Transformer-based models have demonstrated remarkable capabilities in human-like text generation, natural language understanding, and text-to-image [Vaswani *et al.*, 2017; Devlin, 2018]. As transformer-based networks have predominated in various tasks, researchers believe that the transformer architecture can also replicate the success when applied in spiking neural networks.

Zhou et al. proposed a Spiking Transformer model, namely Spikformer, which models images into sparse visual features by using spike-form information without using a softmax operation for the first time [Zhou *et al.*, 2023]. This form of spiking-transformer conducts bio-inspired spatio-temporal dynamics and spike (0/1) activations to obtain high energy efficiency from spiking self-attention. Subsequently, Zhou et al. proposed an SNN-based variant to improve Spikformer to Spikformer-v2 [Zhou *et al.*, 2024b].

The authors developed a Spiking Convolutional Stem (SCS) with supplementary convolutional layers and connected SCS to Spikformer to enhance the image representation by spikes. Additionally, Spikformer-v2 introduced self-supervised learning for directly training SNNs.

Man et al. proposed a Spike-driven Transformer (SDT) [Yao *et al.*, 2024b], which exploited only mask and addition operations without any multiplication and made up to 87.2× lower computation energy than vanilla self-attention. However, the experiment results in this work show that the top-1 accuracy on ImageNet-1K was only 77.1% with 66.34 (M) parameters, which still has a large gap between ANN-based transformers. Therefore, the authors improved SDT to

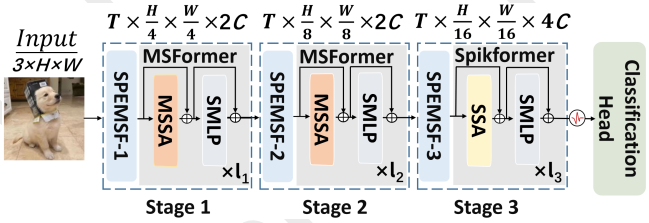

Figure 1: Overview of MSVIT, a hierarchical spiking transformer with multi-scale spiking attention. Note C denotes the spike-form dimension.

SDT-v2 [Yao *et al.*, 2024a], which extended the Spike-driven Transformer into a meta form. On ImageNet-1K, SDT-v2 achieved top-1 accuracy up to 80.0% with 55 (M) parameters, surpassing SDT-v1 by 3.7%. Zhou et al. proposed QKFormer [Zhou *et al.*, 2024a], which enhanced the spiking transformer architecture by using a spiking Q-K attention module. Unlike the comment form of self-attention with Query (Q), Key (K), and Value (V), Q-K attention only adopts Q and K to implement the self-attention mechanism. The authors reported that QKFormer, significantly improved performance compared to Spikformer on ImageNet-1K.

3 Methods

In this section, we first present the overall architecture of MSVIT. Secondly, we introduce the important components of MSVIT, including hybrid spiking attention integration, Spiking Patch Embedding with Multi-scale Feature Fusion (SPEMSF) which serves as the tokenization method for MSVIT, and MSFormer Blocks. Finally, we introduce MSSA and SSA in detail.

3.1 Model Architecture

Overall of Architecture

The overview of MSVIT is illustrated in Figure 1. The input I of MSVIT is represented as $T \times C_0 \times H \times W$, where T denotes the timesteps, C_0 denotes the number of channels, and H and W denote the height and width of the inputs, respectively.

When training MSVIT on static RGB image datasets, T is 1 and C_0 is set to 3. For neuromorphic datasets, $T = T_0$, $C_0 = 2$. In the encoder, A patch size of 4×4 is used, and the input feature $(4 \times 4 \times C_0)$ is transformed into a spike-form representation with C channels (note C distinguishes from C_0) using Spiking Patch Embedding with Multi-scale Feature Fusion-1 (SPEMSF-1).

Given an input $I \in \mathbb{R}^{T \times C_0 \times H \times W}$, SPEMSF-1 first transforms I into a series of tokens x_n $(n \in N)$, where $x \in \mathbb{R}^{T \times H \times W \times C}$ and $N = \frac{H}{4} \times \frac{W}{4}$. Along with the subsequent MSFormer block, this constitutes "Stage 1". To construct a hierarchical spiking transformer, the number of tokens n is further reduced in SPEMSF-2 and SPEMSF-3 to $\frac{H}{8} \times \frac{W}{8}$ and $\frac{H}{16} \times \frac{W}{16}$, respectively. These transformations are handled in "Stage 2" and "Stage 3", where each stage reduces the token dimensions by a 2×2 patch.

In the first and the second stages, the number of channels C is set to 2, while it is increased to 4 in Stage 3. The number of spiking transformer layers L in each stage is configured as l_1, l_2, and l_3, based on the model size and the dataset being trained (detailed settings are provided in Section 4).

Together, the three stages collectively implement the hierarchical spiking-transformer architecture for MSVIT.

Hybrid Spiking Attention Block Integration

While Spiking Neural Networks (SNNs) are inherently energy-efficient, the amount of information that can be processed through spikes in transformer architectures remains limited compared to float values [Zhang *et al.*, 2022; Zhou *et al.*, 2023], resulting in suboptimal performance.

To address this limitation, we integrate multi-scale spiking attention (MSSA) into MSVIT. However, applying a single type of MSSA across all stages of MSVIT leads to an increase in model size, which is not ideal. To break the trade-off between performance and efficiency, we revisit the model design and propose a hybrid spiking attention mechanism for MSVIT. In the first and second stages, we adopt MSSA within the MSFormer block to enhance feature extraction. In the final stage, we employ the standard spiking self-attention (SSA), as used in Spikformer [Zhou *et al.*, 2023], to process the deeper layers of MSVIT. This hybrid design strikes a balance between model size and performance, enabling MSVIT to achieve improved accuracy while maintaining a relatively low energy cost during training.

MSFormer Blocks

Meta-former [Yu *et al.*, 2022] established that the general structure of transformers can be characterized by two key components: a token mixer and a channel mixer. In conventional transformer architectures, the token mixer is often implemented by an attention block, while an MLP block implements the channel mixer.

Similarly, MSVIT adopts this structure and is referred to as MSFormer. Each MSFormer block comprises two modules: a multi-scale spiking attention (MSSA) module and a Spiking MLP (SMLP) module. The formulation of the block can be expressed as follows:

$$X'_l = MSSA(X_{l-1}) + X_{l-1}, X'_l \in \mathbb{R}^{T \times N \times D}, \quad (1)$$

$$X_l = SMLP(X_{l-1}) + X_{l-1}, X_l \in \mathbb{R}^{T \times N \times D}, \quad (2)$$

where N is the length of patches and D is the embedding size.

Spiking Patch Embedding with Multi-scale Feature Fusion

In this section, we describe the proposed Spiking Patch Embedding with Multi-Scale Feature Fusion (SPEMSF) in detail. For vision tasks, higher-level feature maps often preserve rich semantic information but suffer from lower resolution, whereas lower-level feature maps capture raw input details with higher resolution. To address this trade-off, some ANN-based methods [Chen *et al.*, 2018; Badrinarayanan *et al.*, 2017] combine feature maps across layers. However, the improvements in these works are limited due to the semantic gap between feature maps at different scales.

Inspired by these works, we propose a multi-scale feature fusion approach to enhance the spike representation in patch embedding for MSVIT. Our approach leverages convolutional layers with different kernel sizes to process input

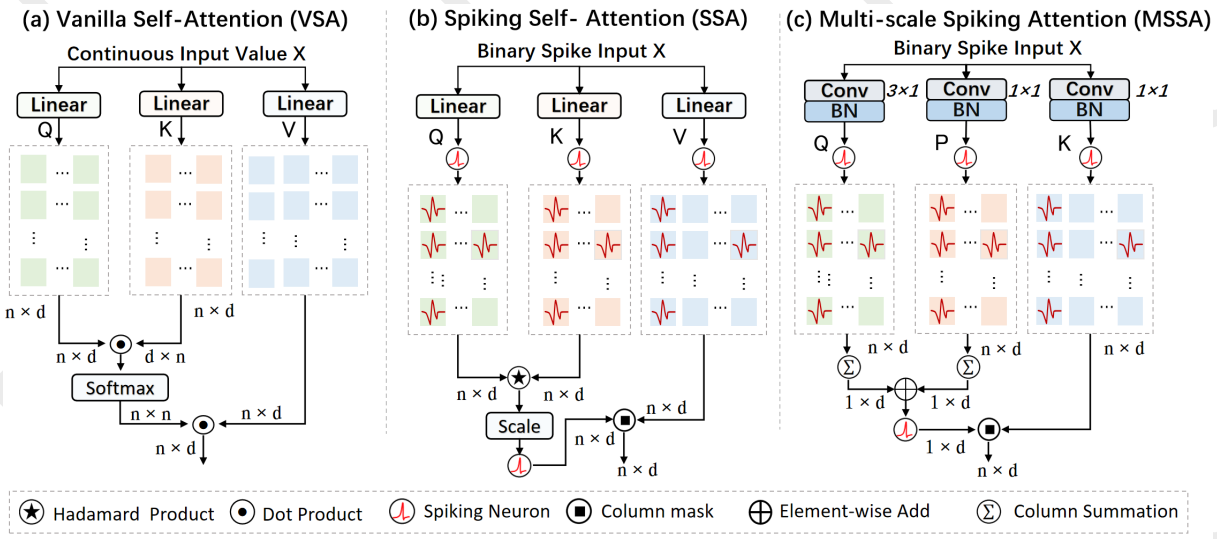

Figure 2: Comparison of three self-attention computation paradigms. (a) VSA employs floating-point matrix multiplication to assess the spatial correlation between Q and K, resulting in a computational complexity of $O(N^2D)$, (b) SSA lacks a dedicated temporal interaction module, maintaining the same complexity as VSA and, (c) MSSA introduces multi-scale interactions, reducing the complexity to $O(T(ND))$.

features at varying semantic levels, effectively treating them as multi-scale features. The fused features, represented as spikes, preserve both low-level and high-level semantic information, reducing information loss during embedding.

The core idea is to use lightweight convolutional operations to project input spiking maps into multiple feature channels. These projections are fused to generate a richer representation for subsequent processing. Specifically, we utilize a combination of linear projections W_d to achieve this.

Given the input spiking map X, the patch embedding process is formulated as:

$$Y = \mathcal{F}(X, W_i) \oplus \mathcal{G}(X, W_j), \tag{3}$$

where $\mathcal{F}$ and $\mathcal{G}$ are functions representing different multi-scale transformations, and $\oplus$ denotes the fusion operation.

In our implementation:

- The linear projection W_d in function $\mathcal{F}$ is defined as a lightweight convolutional layer with a 1×1 kernel and stride ≥ 1, which focuses on channel-wise transformations,
- function $\mathcal{G}$ uses a 3×3 convolutional layer with stride $= 2$, incorporating more spatial context while reducing resolution.

The implementation for function $\mathcal{F}$ is a simple pipeline of Conv2D-BN-SNN, while $\mathcal{G}$ is designed with one of the following configurations: (1) Conv2D-BN-MaxPooling-SN-Conv2D-BN-SNN, or (2) Conv2D-BN-SN-Conv2D-BN-MaxPooling-SNN. This multi-scale feature fusion not only enriches the spike representation for inputs but also ensures compatibility with the channel and token requirements of the patch embedding block. The figure of Spiking Patch Embedding is further illustrated in Appendix B.

3.2 Multi-scale Spiking Attention (MSSA)

We propose a novel Spike-driven Transformer that maintains the spike-driven nature of Spiking Neural Networks (SNNs) throughout the network while achieving strong task performance by incorporating multi-scale features into the attention module.

The overview of Multi-Scale Spiking Attention (MSSA) is presented in Figure 2 (c). For comparison, traditional Vanilla Self-Attention (VSA) and Spiking Self-Attention (SSA) which is the core component of Spikformer, are shown in Figure. 2 (a) and (b). Both VSA and SSA use three components (Q, K, V) and have computational complexity of $O(N^2d)$ or $O(Nd^2)$. In contrast, MSSA achieves linear complexity of $O(Nd)$. The initialization of the three components in MSSA is defined as follows:

$$Q = \mathcal{SN}(BN(XW_Q)), \tag{4}$$
$$P = \mathcal{SN}(BN(XW_P)), \tag{5}$$
$$V = \mathcal{SN}(BN(XW_V)), \tag{6}$$

where $X \in \{0,1\}^{T\times N\times D}$ is the spiking map of the input, N represents the number of patches, and D is the feature dimension. Q and P represent lower and higher-level spiking feature maps of X, generated by convolutional layers. All three components (Q, P, V) are produced through learnable linear matrices. Here, $\mathcal{SN}$ denotes the spiking neuron layer, and BN represents the batch normalization layer.

Unlike VSA (a) and SSA (b), MSSA replaces matrix multiplication with column summation to compute the attention interaction between the components (Q, P, V). The attention mechanism in MSSA is defined as:

$$\text{MSSA}(Q, P, V) = \mathcal{SN}(\text{SUM}_c(Q) \oplus \text{SUM}_c(P)) \otimes V, \tag{7}$$

where $\oplus$ represents element-wise addition, and $\text{SUM}_c(\cdot)$ performs column-wise summation, resulting in an $N \times 1$ vector

α, which encodes self-attention scores in spike form:

$$\alpha_q = \sum_{i=0}^{d} Q_i, \quad \alpha_p = \sum_{i=0}^{d} P_i. \tag{8}$$

To implement multi-scale feature fusion, α_q and α_p are mixed by element-wise addition ($\oplus$), fusing the lower-level feature Q with the higher-level feature P via the attention scores. The resulting spike-based attention vector is then applied to the V spike matrix to generate the final spike representation of X using MSSA.

3.3 Spiking Self Attention

Spikformer [Zhou *et al.*, 2023] introduced a novel spike-based self-attention mechanism, termed Spiking Self-Attention (SSA). Unlike traditional self-attention, SSA leverages sparse spike-form representations (Q, K, V) and eliminates the need for softmax operations and floating-point matrix multiplications. The computational process of SSA is formulated as:

$$Q = \mathcal{SN}(BN(XW_Q)), \tag{9}$$
$$K = \mathcal{SN}(BN(XW_K)), \tag{10}$$
$$V = \mathcal{SN}(BN(XW_V)), \tag{11}$$
$$SSA(Q, K, V) = \mathcal{SN}((QK^T) \otimes V \cdot s), \tag{12}$$

where $Q, K, V \in \mathbb{R}^{T \times N \times D}$ are spike-form representations computed through learnable linear transformations. Here, s is a scaling factor. $\mathcal{SN}$ denotes the spiking neuron layer and BN represents the batch normalization layer.

In stage 3 of MSVIT, we adopt SSA to perform spiking attention in the deeper layers, leveraging its efficiency and alignment with the spike-driven computation paradigm.

4 Experiments

This section introduces the details of the experiment, including data collection, implementation, and evaluation methods.

4.1 Experimental Setup

We evaluate MSVIT on both static image classification and neuromorphic classification tasks. For static image classification, we use ImageNet-1K [Deng *et al.*, 2009] and CIFAR10/100 [Krizhevsky *et al.*, 2009]. For neuromorphic classification, we employ the CIFAR10-DVS [Li *et al.*, 2017] and DVS128 Gesture [Amir *et al.*, 2017] datasets. Appendix D introduces the datasets in detail.

4.2 Results on ImageNet-1K Classification

Experimental Setup on ImageNet-1K. We adopt a training recipe similar to that proposed in [Zhou *et al.*, 2024a] and detail the configurations in this section. First, the model is trained in a distributed manner for 200 epochs on an 8-A100 GPU server. We employ several data augmentation techniques, including RandAugment [Cubuk *et al.*, 2020], random erasing [Zhong *et al.*, 2020], and stochastic depth [Huang *et al.*, 2016], with a batch size of 512. Additionally, gradient accumulation is utilized to stabilize training, as suggested in [He *et al.*, 2022]. Second, the optimization process leverages synchronized AdamW with a base learning rate of 6×10^{-4} per batch size of 512. The learning rate is linearly warmed up at the initial stage and subsequently decays following a half-period cosine schedule. The effective runtime learning rate is scaled proportionally to the batch size, calculated as `BatchSize/256` multiplied by the base learning rate. Finally, the architecture is designed with three stages (as illustrated in Figure 1), where the number of layers in each stage is configured as $\{l_1 = 1, l_2 = 2, l_3 = 7\}$, respectively. These configurations collectively ensure robust and efficient training of the proposed model.

Primary Results on ImageNet-1K. The experimental results demonstrate the superior performance of our proposed MSVIT, surpassing previous works' performance. Generally, MSVIT (69.80 M) with input size of 224^2 achieves 85.06% top-1 accuracy and 97.58% top-5 accuracy on ImageNet-1K, which performs the best in Table 1. To start with the experiments, we first compare MSVIT with Spikformer which is the first version of the spike-form transformer[Yao *et al.*, 2022]. Our MSVIT (69.80 M, 85.06%) significantly outperforms Spikformer (66.34 M, 74.81%, by 10.25% with the input size of 224^2. In Table 1, we can see that MSVIT achieves the best performance on accuracy, utilizing slightly more parameters. Additionally, compared to SDSA, MSSA has lower computational complexity. Meanwhile, MSVIT outperforms Spike-driven Transformer [Yao *et al.*, 2024b] (SDT, built by SDSA) by 7.81%, 8.39%, and 8.58% respectively at three model size levels (17.69M, 30.23M, 69.80M), and surpasses QK-Former by 1.29%, 0.92%, and 0.84 % by comparable parameters. Surprisingly, MSVIT obtains significant improvement, outperforming MST-T [Wang *et al.*, 2023b] by 4.45% with 30.23M parameters, and 45.88 mJ energy cost.

Comparing with ANN Models on ImageNet. MSVIT is designed as an event-driven SNN model, in which the outputs of the embedding layers, the matrix calculation in the attention blocks, and information transmission are binary spikes $\{0, 1\}$. As a result, the multiplications of the weight matrix and activations, such as RELUs, which are important, can be replaced by AND-ACcumulate (AC) operations, which benefit model training with high energy efficiency. Meanwhile, to achieve a competitive performance for ANN-based models, we apply the hierarchical transformer architecture to MSVIT. Although the implementation of hierarchical architectures is relatively more complex than those using the same attention blocks throughout the networks ([Zhou *et al.*, 2023; Yao *et al.*, 2024b]), it is still worth using a hierarchical architecture to reduce the performance gap between ANNs and SNNs. This is mainly because the hierarchical architecture naturally has the flexibility to model at various scales and has linear computational complexity with respect to image input size [Liu *et al.*, 2021]. For instance, our MSVIT outperforms the most well-known Transformer-based ANNs in performance with high energy efficiency under the same experiment conditions without pretraining or extra training data, among MSVIT (69.80M, 85.06%, SNN, 45.88mJ), Swin Transformer (88M, 84.5%, ANN, 216.20mJ) [Liu *et al.*,

Method	spiking	Architecture	Params (M)	Input Size	Time Step	Energy (mJ)	Top-1 Acc. (%)
DeiT [Touvron *et al.*, 2021]	✗	DeiT-B	86.60	224^2	1	80.50	81.80
VIT-B/16 [Dosovitskiy, 2020]	✗	ViT-12-768	86.59	384^2	1	254.84	77.90
Swin Transformer [Liu *et al.*, 2021]	✓	Swin-T	28.50	224^2	1	70.84	81.35
	✗	Swin-S	51.00	224^2	1	216.20	83.03
Spikformer [Zhou *et al.*, 2023]	✓	8-384	16.80	224^2	4	5.97	70.24
	✓	8-768	66.30	224^2	4	20.0	74.81
Spikformer V2 [Zhou *et al.*, 2024b]	✓	V2-8-384	29.11	224^2	4	4.69	78.80
	✓	V2-8-512	51.55	224^2	4	9.36	80.38
Spike-driven [Yao *et al.*, 2022]	✓	SDT 8-384	16.81	224^2	4	3.90	72.28
	✓	SDT8-512	29.68	224^2	4	4.50	74.57
	✓	SDT8-768	66.34	224^2	4	6.09	77.07
Spike-driven v2 [Yao *et al.*, 2024a]	✓	SDT v2-10-384	15.10	224^2	4	16.70	74.10
	✓	SDT v2-10-512	31.30	224^2	4	32.80	77.20
	✓	SDT v2-10-768	55.40	224^2	4	52.40	80.00
QKFormer [Zhou *et al.*, 2024a]	✓	QK-10-384	16.47	224^2	4	15.13	78.80
	✓	QK-10-512	29.08	224^2	4	21.99	82.04
	✓	QK-10-768	64.96	224^2	4	38.91	84.22
MSVIT	✓	MSVIT-10-384	17.69	224^2	4	16.65	**80.09**
	✓	MSVIT-10-512	30.23	224^2	4	24.74	**82.96**
	✓	MSVIT-10-768	69.80	224^2	4	45.88	**85.06**

Table 1: Comparison of MSVIT performance to respective ANN-based and SNN-based state-of-the-art models by accuracy (Acc.%). The experimental results are on ImageNet-1K. Energy is calculated as the average theoretical power consumption when predicting an image from ImageNet test set. The energy data for MSVIT and ANNs is evaluated according to Appendix C.

2021], DeiT-B (86M, 83.1%, ANN, 254.84mJ) [Touvron *et al.*, 2021] and ViT (85.59M, 77.9%, ANN, 254.84mJ) [Dosovitskiy, 2020] (the detailed results are in Table 1).

4.3 Results on CIFAR and Neuromorphic Datasets

CIFAR Classification. We conduct experiments on smaller datasets and configure the training process to ensure sufficient model optimization. Specifically, we train the model for 400 epochs with a batch size of 128, following the setup of Spikformer [Zhou *et al.*, 2023]. For the three-stage training of MSVIT, we utilize a total of 4 blocks distributed as $\{1, 1, 2\}$ across the stages. Thanks to the hierarchical architectural design, MSVIT comprises 7.59M parameters, which is slightly larger than QKFormer (6.74M) but smaller than Spikformer (9.32M). The performance results on the CIFAR datasets are summarized in Table 2. On CIFAR10, MSVIT achieves an accuracy of **96.53%**, outperforming Spikformer by 1.02% and QKFormer [Zhou *et al.*, 2024a] by 0.35%. For CIFAR100, MSVIT achieves an accuracy of **81.98%**, exceeding Spikformer (78.21%) by **3.77%** and QKFormer by 0.86%. Notably, MSVIT surpasses Vision Transformer (ViT), an ANN-based model, on CIFAR100 by 0.93%. This represents a relatively significant improvement over other SNN-based Transformer architectures, demonstrating the efficacy of MSVIT in classification tasks on various datasets.

Neuromorphic Classification. To evaluate MSVIT on neuromorphic tasks, we compare our model with the state-of-the-art models using the CIFAR10-DVS and DVS-Gesture datasets. Unlike conventional static image datasets, neuromorphic datasets comprise event streams instead of RGB images. This introduces a significant domain shift between the source (static image datasets) and target (neuromorphic datasets) domains for models pre-trained on static images. We aggregate events over specific time intervals to address this discrepancy to construct frames. The RGB channels are replaced with positive, negative, and the sum of events as input features, respectively. For this experiment, we implement a lightweight version of MSVIT with only 1.67M parameters, utilizing a block configuration of $\{0, 1, 1\}$ across the three stages. The maximum patch embedding dimension is set to 256. The model is trained for 200 epochs on the DVS128-Gesture dataset and 106 epochs on the CIFAR10-DVS dataset. The number of time steps for the spiking neurons is set to either 10 or 16. The experimental results for temporal neuromorphic classification are summarized in Table 2. On the DVS128-Gesture dataset, MSVIT with 1.67M parameters achieves an accuracy of 98.80% using 16-time steps and 98.37% using 10-time steps. For the CIFAR10-DVS dataset, MSVIT achieves an accuracy of 84.30% using 16-time steps, significantly outperforming Spikformer by 3.4%. Moreover, with 10-time steps, MSVIT achieves an accuracy of 83.80%, surpassing Spikformer by 4.20% and QKFormer by 0.3%. These results highlight the efficiency and performance gains of MSVIT, with a minimal parameter count.

method	CIFAR10			CIFAR100			DVS128			CIFAR10-DVS		
	Param	T	Acc	Param	T	Acc	Param	T	Acc	Param	T	Acc
Spikformer	9.32	4	95.51	9.32	4	78.21	2.57	16	98.3	2.57	16	80.9
SDT [Yao *et al.*, 2024b]	9.32	4	95.81	9.32	4	78.21	2.57	16	99.3	2.57	16	80.9
CML [Wang *et al.*, 2023a]	9.32	4	95.81	9.32	4	80.02	2.57	16	98.6	2.57	16	80.9
QKFormer [Zhou *et al.*, 2024a]	6.74	4	96.08	6.74	4	81.12	1.50	16	98.6	1.50	16	84.0
ResNet-19	12.63	-	94.97	12.63	-	75.35	-	-	-	-	-	-
Transfomer (4-384)	9.32	-	**96.73**	9.32	-	81.02	-	-	-	-	-	-
MSVIT	7.59	4	96.53	7.59	4	**81.98**	1.67	16	98.80	1.67	16	**84.30**

Table 2: Comparision on CIFAR10, CIFAR100, DVS128, CIFAR10-DVS. "Param" denotes "Parameter (M)", "Acc" denotes "Top-1 Accuracy (%)", "T" denotes "Time Step".

Model	Param (M)	CIFAR100 (Acc)
MSSA (P+P) +SSA	7.74	81.36
MSSA (Q+Q) +SSA	7.45	81.56
MSSA (P) +SSA	7.52	81.35
MSSA(Q) +SSA	7.37	81.44
MSSA +MSSA(P+P)	9.04	81.25
MSSA +MSSA (Q+Q)	7.89	81.15
MSSA +MSSA (P+Q)	8.48	81.65
MSSA(P+Q) + SSA (MSVIT)	7.59	**81.98**

Table 3: Ablation study of MSSA with different feature fusion dimensions. P: feature from 3×1 conv; Q: feature from 1×1 conv.

5 Ablation Study

Hybrid Spiking Attention Integration. We test MSVIT on CIFAR100 and use the MSVIT equipped with MSSA (on stage 1,2) and SSA (on stage 3) as the baseline. The results show that using the same MSSA at each stage achieves relatively high performance with 81.65% on Top-1 accuracy among the study cases in Table 3. However, the number of parameters of MSVIT increases too much, which may incur Unworthy computational consumption, especially in the case we conduct experiments on large-scale datasets, such as the ImageNet-1K. This is mainly because a (3×3) kernel in convolution layers consumes more parameters than a pointwise convolution layer (with (1×1) kernels) to extract features from a larger perception field. Most of the baselines, including Spikformer [Zhou *et al.*, 2023], QKFormer [Zhou *et al.*, 2024a], and SDT [Yao *et al.*, 2024a], solely use the PointWise Convolution layer (PWConv) to conduct information extractions. Although PWConv decreases the number of parameters throughout the model, it may cause information loss of high-level features at the shallow layers of spike-form transformer architectures, which are stage 1 and stage 2 in our MSVIT. Particularly, using the binary spikes that only use $\{0, 1\}$ to transfer information is difficult. We addressed this issue and conducted the fusion of the low-level feature Q, and the high-level feature P to transmit more features to the model. Finally, we adopted the hybrid spiking attentions (MSSA + SSA) as the final version of MSVIT to strike the trade-off between the computational efficiency and performance of the model. The experimental results also show that MSVIT gains great performance with 81.98% on CIFAR100 using only 7.59M parameters.

The effectiveness of MSSA. Lines 1- 4 in Table 3 illustrate the effectiveness of feature fusion at stages 1 and 2 of MSVIT. Line 1 shows that MSSA (P+P) performs 81.25, which is almost the same as the result of MSSA (P) (Line 3). This indicates that the fusion for the same feature sizes can not improve its performance for MSVIT. Line 2 shows MSSA(Q+Q) obtains only 81.15, which is similar to QKformer's (only using 1×1 kernels) experimental result. In conclusion, adopting only low-level/local feature maps may still incur semantic information loss, damaging the spiking representation of input images.

6 Conclusion

In this work, we design a novel spike-driven multi-scale attention (MSSA), which involves a multi-scale feature fusion mechanism in the spiking hierarchical transformer architecture to improve the model's performance. MSSA fuses low-level and high-level information of inputs, significantly improving the model performance with limited parameter increases. Furthermore, MSSA replaces the dot-product or Hadamard product existing in VSA or SSA with a column sum, maintaining the computation in linear complexity to the tokens of the inputs. Correspondingly, we also utilize Spiking Patch Embedding with Multi-scale Feature Fusion (SPEMSF), which enhances spiking representation for both high-level and low-level information of inputs, thereby promoting model improvement. Finally, we implement a hierarchical spiking transformer architecture equipped with the aforementioned MMSA and SPEMSF, namely MSVIT. We have conducted extensive experiments, and the results show that our model achieves state-of-the-art performance on both static and neuromorphic datasets. Notably, MSVIT achieved top-1 accuracy on ImageNet-1K over 85% with 69.80M parameters and image input of size 224^2 by direct training from scratch. Leveraging the MSVIT's superior capabilities, we strive to inspire confidence in the application of Spiking Neural Networks through our work.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YDLN0011, 2024YDLN0006) and the Research Grants Council of the Hong Kong SAR (Grant No. PolyU25216423).

References

[Amir *et al.*, 2017] 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.

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[Cubuk *et al.*, 2020] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[Deng *et al.*, 2009] 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. Ieee, 2009.

[Devlin, 2018] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fan *et al.*, 2021] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[Graham *et al.*, 2021] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021.

[Guo *et al.*, 2020] Qipeng Guo, Xipeng Qiu, Pengfei Liu, Xiangyang Xue, and Zheng Zhang. Multi-scale self-attention for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7847–7854, 2020.

[Guo *et al.*, 2021] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Huang *et al.*, 2016] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016.

[Jie and Deng, 2023] Shibo Jie and Zhi-Hong Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1060–1068, 2023.

[Koenderink, 1984] Jan J Koenderink. The structure of images. *Biological cybernetics*, 50(5):363–370, 1984.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Li *et al.*, 2017] 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.

[Li *et al.*, 2023] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16889–16900, 2023.

[Liu *et al.*, 2021] 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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Jing Liu, Zizheng Pan, Haoyu He, Jianfei Cai, and Bohan Zhuang. Ecoformer: Energy-saving attention with linear complexity. *Advances in Neural Information Processing Systems*, 35:10295–10308, 2022.

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

[Mehta and Rastegari, 2021] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose,

and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[Mei *et al.*, 2021] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3517–3526, 2021.

[Neil and Dirk, 2020] Houlsby Neil and Weissenborn Dirk. Transformers for image recognition at scale. *Online: https://ai. googleblog. com/2020/12/transformers-for-image-recognitionat. html*, 2020.

[Pan *et al.*, 2020] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020.

[Park *et al.*, 2019] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[Pei *et al.*, 2019] Jing Pei, Lei Deng, Sen Song, Mingguo Zhao, Youhui Zhang, Shuang Wu, Guanrui Wang, Zhe Zou, Zhenzhi Wu, Wei He, et al. Towards artificial general intelligence with hybrid tianjic chip architecture. *Nature*, 572(7767):106–111, 2019.

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

[Shi *et al.*, 2024] Xinyu Shi, Zecheng Hao, and Zhaofei Yu. Spikingresformer: Bridging resnet and vision transformer in spiking neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2024.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[Wang *et al.*, 2023a] 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 the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, volume 8, pages 3085–3093, 2023.

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

[Whittington *et al.*, 2018] James Whittington, Timothy Muller, Shirely Mark, Caswell Barry, and Tim Behrens. Generalisation of structural knowledge in the hippocampal-entorhinal system. *Advances in neural information processing systems*, 31, 2018.

[Yao *et al.*, 2022] 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.

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

[Yao *et al.*, 2024b] 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.

[Yu *et al.*, 2022] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.

[Zhang *et al.*, 2022] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8801–8810, 2022.

[Zhong *et al.*, 2020] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

[Zhou *et al.*, 2023] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *ICLR*, 2023.

[Zhou *et al.*, 2024a] Chenlin Zhou, Han Zhang, Zhaokun Zhou, Liutao Yu, Liwei Huang, Xiaopeng Fan, Li Yuan, Zhengyu Ma, Huihui Zhou, and Yonghong Tian. Qkformer: Hierarchical spiking transformer using qk attention. *CoRR*, 2024.

[Zhou *et al.*, 2024b] Zhaokun Zhou, Kaiwei Che, Wei Fang, Keyu Tian, Yuesheng Zhu, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer v2: Join the high accuracy club on imagenet with an snn ticket. *CoRR*, 2024.