Tackling Long-Tailed Data Challenges in Spiking Neural Networks via Heterogeneous Knowledge Distillation

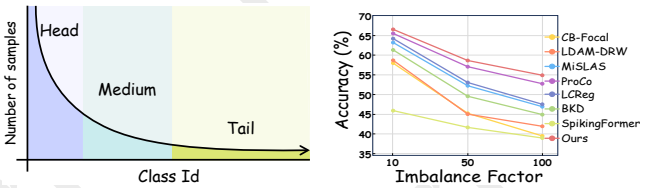
Moqi Li , **Xu Yang** , **Cheng Deng***
Xidian University
{moqili14, xuyang.xd, chdeng.xd}@gmail.com

Abstract

Spiking Neural Networks (SNNs), inspired by the behavior of biological neurons, have gained significant research interest for resource-constrained edge devices and neuromorphic hardware due to the usage of inter-unit communication binary spike signals with low power consumption. However, the absence of research on spiking neural networks on long-tailed data has severely limited the deployment and application of this emerging network in practical scenarios. To fill this gap, this paper proposes a long-tailed learning framework based on spiking neural networks, named LT-SpikingFormer, to alleviate the distribution bias between head and tail classes. LT-SpikingFormer adopts a widely trained Convolutional Neural Network to construct a heterogeneous knowledge distillation paradigm, offering balanced and reliable prior knowledge. Moreover, a multi-granularity hierarchical feature distillation objective is proposed for cross-layer local features and network global predictions to facilitate refined information distillation to optimize the network, specifically for the performance of the tailed classes. Extensive experimental results demonstrate that our method performs well on several benchmark datasets.

1 Introduction

Spiking Neural Networks (SNNs) [Maass, 1997], modeled after biological neurons, have garnered considerable research attention because they use binary spike signals for communication between units. Unlike conventional Artificial Neural Networks (ANNs), which rely on continuous activation functions, spiking neurons encode continuous input values into spike trains using models like the Leaky Integrate-and-Fire (LIF) neuron model and its variants, such as the PLIF. This spike-based computation allows for information encoding in both spatial and temporal domains, potentially improving computational efficiency and reducing energy consumption. As a result, SNNs [Zhao *et al.*, 2024; Shen *et al.*, 2025b; Shen *et al.*, 2025a] are well-suited for real-time processing tasks and neuromorphic hardware applications.



(a) Long-tailed distribution. (b) Top-1 Accuracy.

Figure 1: The label distribution of long-tailed dataset and the top-1 accuracy of current mainstream methods in CIFAR10-LT. (a) The label distribution map shows that the head class feature space learned from these samples is usually larger than the tail class feature space. (b) The line chart provides an intuitive visualization of how dataset imbalance influences model performance.

Two primary methodologies for constructing spiking neural networks with standard spiking behavior exist. The first approach is ANN-to-SNN conversion [Diehl *et al.*, 2015; Sengupta *et al.*, 2019], where spiking neurons replace ReLU activation layers in ANNs. This conversion method, however, often necessitates a large number of time steps to approximate the ReLU function accurately, leading to increased latency. The second approach, direct training of SNNs [Mostafa, 2017; Neftci *et al.*, 2019], a widely utilized method, involves unfolding the network across simulation time steps and applying backpropagation through time. Due to the non-differentiable nature of the spike-generation process, this backpropagation typically relies on surrogate gradients. While numerous ANN models have been successfully adapted to SNNs, their application to long-tailed data distributions, which are common in real-world tasks, remains underexplored.

Long-tailed distribution is prevalent in real-world applications. As shown in Fig. 1a, such datasets exhibit highly imbalanced sample distributions, posing significant challenges. While numerous approaches have been developed to address this problem in ANNs, little exploration has been done into the promising, efficient, and robust SNN architecture. This absence of research has significantly hindered the deployment and effectiveness of SNNs in practical scenarios. Moreover, we found that methods that aim to solve the long-tailed problem faced by ANNs are challenging to adapt to the unique

*Corresponding Author

architecture of SNNs. To address this limitation, our paper introduces an efficient knowledge distillation framework designed to optimize SNN performance in long-tailed data settings, providing a straightforward and effective approach for high-performance model adaptation.

This work proposes a novel direct training framework for SNNs that handles diverse, long-tailed datasets. The proposed framework, LT-SpikingFormer, integrates a self-attention mechanism to enable dynamic interactions among spiking features. To bridge the impact of long-tailed distribution on the model, we construct a new CNN-SNN heterogeneous knowledge distillation paradigm to extract robust knowledge from widely trained CNNs to alleviate overfitting in most categories and enhance balanced feature learning across categories. Toward this heterogeneous knowledge distillation optimization, we propose a multi-granularity hierarchical feature distillation objective that leverages cross-layer local features and network global predictions to facilitate refined information distillation to optimize the network, specifically for the performance of the tailed classes. Extensive experimental results [Cui *et al.*, 2019; Liu *et al.*, 2024; Wang *et al.*, 2024; Ma *et al.*, 2024] demonstrate that our method can achieve state-of-the-art results, as shown in Fig. 1b.

Contributions. The highlights of the paper are threefold: 1) By analyzing and summarizing the existing work, we propose the first spiking neural network for long-tailed data and employ the heterogeneous distillation paradigm of CNN-SNN to alleviate the imbalance of samples, filling the research gap of long-tailed spiking learning; 2) For the proposed heterogeneous distillation framework, we constructed a joint optimization manner consisting of global and local distillations, in which the local one with norm guided can effectively alleviate the long-tailed distribution in different latent space, while the global knowledge distillation based on finally predictions can significantly improve the overall recognition performance; 3) Extensive experimental results demonstrate that our method achieves state-of-the-art performance on several datasets. The ablation experiments are conducted to verify the effectiveness of each module.

2 Related Work

Long-tailed Visual Recognition. Long-tailed distribution is a significant challenge in machine learning and visual recognition. In image classification, common object categories dominate the dataset, while rare categories are underrepresented, leading to model bias toward the "head" categories during training, which causes overfitting, while generalization to the "tail" categories with fewer samples is limited, degrading overall performance.

Traditional methods to mitigate the long-tailed problem involve resampling and reweighting. Resampling methods aim to balance sample sizes across categories by oversampling tail categories [Li *et al.*, 2022] or undersampling head categories [He and Garcia, 2009], though these can lead to overfitting or underfitting. Reweighting [Khan *et al.*, 2017] adjusts the learning rates for categories based on sample sizes, though such methods can improve performance on tail categories at the expense of head categories. Two-stage training has been proposed to address this: the first stage focuses on feature representation learning on the original long-tailed distribution, while the second applies resampling or reweighting for fine-tuning, optimizing both feature learning and classifier learning.

Transfer learning [Long *et al.*, 2022] and knowledge distillation [Zhao *et al.*, 2023; Jin *et al.*, 2023] also offer solutions. Transfer learning facilitates knowledge transfer from majority classes to minority classes, while knowledge distillation uses soft labels from a large teacher model to transfer knowledge to a smaller student model. Recent methods like the mixture of experts (MoE) model, an ensemble approach, assigns experts to focus on different categories or dynamically adjusts based on sample sizes.

Spiking Vision Transformers. Spikformer [Zhou *et al.*, 2022] is the first hybrid architecture to integrate spiking neural networks with Transformers. It introduces a spiking self-attention mechanism that eliminates traditional multiplication operations by activating the query, key, and value with spiking neurons and replacing the softmax function with spiking neurons. Furthermore, it substitutes the layer normalization and GELU activation used in Transformers with batch normalization and spiking neurons. The Spike-driven Transformer [Yao *et al.*, 2024] proposes a linear-complex peak-driven self-attention mechanism designed to improve spatiotemporal information processing and significantly reduce energy consumption. SpikingResformer [Shi *et al.*, 2024] is a spiking neural network architecture that combines the strengths of ResNet and Vision Transformer, enhancing performance while reducing parameter count and energy consumption through the introduction of a double spiking self-attention mechanism.

Knowledge Distillation. Knowledge distillation (KD) enhances the performance of a student model by guiding it to "imitate" the more complex and higher-performing teacher model. In KD, the soft labels produced by the teacher model serve as the target to transfer knowledge from the large teacher model to the smaller student model. There are two primary approaches to knowledge distillation: logit-based KD and feature-based KD. Logit-based KD methods [Cho and Hariharan, 2019; Chen *et al.*, 2024] leverage the output of the teacher model as supervision to direct the learning of the student model, focusing on mimicking the teacher's decision-making process. In contrast, feature-based KD methods [Ahn *et al.*, 2019; Tung and Mori, 2019; Liang *et al.*, 2024] aim to align teacher and student models at the feature representation level by minimizing the distance or discrepancy between their intermediate feature layers. More recently, the Heterogeneous Bridge Distillation [Hao *et al.*, 2023] method has been introduced, which addresses the disparity between heterogeneous features by projecting the intermediate features of the student model into a latent space that aligns with the output of the teacher model.

3 Methodology

This section proposes a novel framework, LT-SpikingFormer, which trains Spiking Neural Networks (SNNs) from scratch

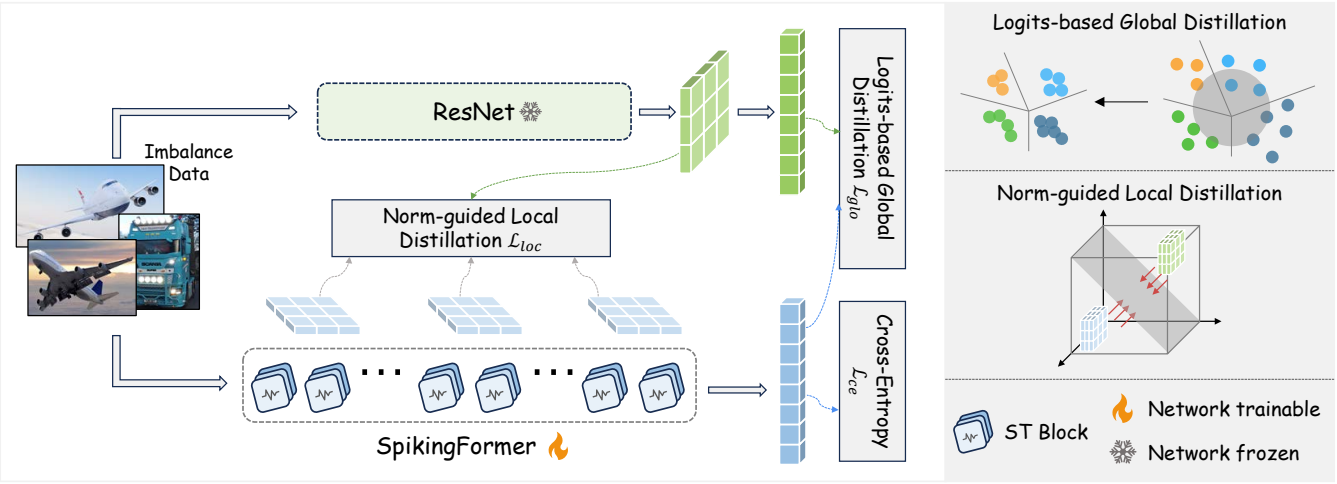


Figure 2: The framework of LT-SpikingFormer.

with a heterogeneous distillation learning paradigm, addressing the challenges of long-tailed data distributions. LT-SpikingFormer adopts a compact ResNet-32 model with minimal augmentation as the teacher to explore the synergistic potential of combining SNNs and CNNs and enhance the recognition capabilities of minority classes through distinctive spiking characteristics. Considering the massive difference between convolutional and spiking networks, we construct a heterogeneous distillation optimization strategy from local to global modules, which is particularly important for fine-grained recognition of underrepresented categories in long-tailed datasets. The framework is shown in Fig. 2.

3.1 Framework of LT-SpikingFormer

As the basic unit of the SNN, the spiking neuron receives the generated current and accumulates the membrane potential to compare with the threshold to determine whether to create a spike. We describe the dynamics of the LIF (Leaky Integrate-and-Fire) neuron by the following discrete-time model:

$$U[t] = V[t-1] + \frac{1}{\tau}(X[t] - (V[t-1] - V_{reset})), \quad (1)$$

$$S[t] = H(U[t] - V_{th}), \quad (2)$$

$$V[t] = U[t](1 - S[t]) + V_{reset}S[t], \quad (3)$$

where $U[t]$ and $V[t]$ are the membrane potentials of neurons before and after charging, $X[t]$ is the input current at time step t, and τ is the membrane time constant. Eq. (2) describes the firing process, where $H(\cdot)$ is the Heaviside step function. When the membrane potential $U[t]$ exceeds the firing threshold V_{th}, the spike neuron will trigger a spike $S[t]$. Eq. (3) describes the resetting process. The membrane potential $V[t]$ after the trigger event is equal to $U[t]$ if no spike is generated, otherwise it is equal to the reset potential V_{reset}.

The Spike Neurons Layer comprises multiple LIF neurons, a key component for information encoding and transmission in SNN. In SNN, the input signal enters the SN layer for processing, where each LIF neuron decides when to emit a pulse based on the input current and its dynamic characteristics (such as membrane time constant, threshold, etc.). These pulses can then be further propagated and processed in the SNN to complete complex computational tasks. For simplicity and clarity in subsequent chapters, we use $\mathcal{SN}(\cdot)$ to represent the spiking neuron layer, omitting the dynamic process inside neurons.

Given 2D images $I \in \mathbb{R}^{N\times H\times W\times 3}$, a Spiking Feature Pre-Extractor $\mathcal{G}(\cdot)$ is used to extract local features $X = [x_1, x_2, \cdots, x_N] \in \mathbb{R}^{N\times T\times D}$, where T, N, H, W, D denote time step, sample number, height, weight, and embedding dimension, respectively. We aim to learn a set of mappings $X \to Y = [y_1, y_2, \cdots, y_N] \in \mathbb{R}^{N\times C}$ that enables accurate sample classification, where C is the number of classes.

$$X = \mathcal{G}(I) \quad , X \in \mathbb{R}^{N\times T\times D}. \quad (4)$$

It is crucial to adopt strategies that ensure the model's robustness and generalization to address the challenges of long-tailed data distributions. An approach is to incorporate a balanced sampling technique alongside advanced image enhancement methods. Specifically, we utilize the imbalanced data sampler strategy to ensure adequate representation of minority classes during training. This sampler assigns sample weights based on the effective number of samples per class, promoting balanced class exposure and mitigating the risks of overfitting and under-sampling. In addition to conventional data augmentation strategies, such as rotation, flipping, and cropping, we also employ advanced methods like Mixup and CutMix, which have been proven to be very effective in alleviating the data imbalance problem.

On the other hand, to fully explore the association relationship in long-tailed data using Spike Neurons, we stack multiple Spiking Transformer (ST) blocks and represent them as $\mathcal{F}(\cdot)$. As illustrated in Fig. 3, each ST block is composed of a Spike-based Self-Attention (SSA) block and a Spike MLP (SMLP) block, and residual connections are applied to each SSA block and SMLP block. The ST block can be formulated as follows:

$$\hat{X}^i = \text{SSA}(X^{i-1}) + X^{i-1}, \quad (5)$$

$$X^i = \text{SMLP}(\hat{X}^i) + \hat{X}^i, \quad (6)$$

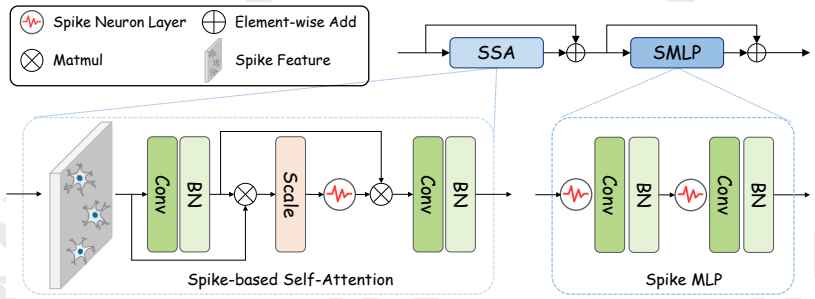


Figure 3: The framework of Spiking Transformer blocks. The overview of Spiking Transformer blocks, which consists of a Self-Attention Block and a Spike MLP Block.

where X^i is the output feature of i-th ST block. The SMLP consists of two feed-forward layers (FFL), and $\text{FFL}(\cdot) = \text{BN}(\text{Conv}(\mathcal{SN}(\cdot)))$.

For the Spike-based Self-Attention (SSA), we introduce a fully spike-driven self-attention mechanism, which is defined as follows:

$$\text{Attn}(X^i) = \mathcal{SN}(\mathbb{E}^\top(X^i, X^i; \text{BN}(\text{Conv}(X^i))) * \sigma_1), \quad (7)$$

$$\text{SSA}(X^i) = \mathcal{SN}(\mathbb{E}(\text{Attn}(X^i), X^i; \text{BN}(\text{Conv}(X^i))) * \sigma_2), \quad (8)$$

where $\mathbb{E}$ is a linear dual-spike transformation function. σ_1 and σ_2 are the scaling factors. The key lies in the design of the scaling factors σ_1 and σ_2, which are adjusted based on the average firing rates of the input features and the attention map.

Finally, we perform global average pooling (GAP) and full connection classification (FC) on the features processed by the stacked Spiking Transformer encoders to obtain the final output prediction Y. SpikingFormer is defined as follows:

$$Y = \text{FC}(\text{GAP}(\mathcal{F}(X))). \quad (9)$$

3.2 Heterogeneous Knowledge Distillation

To harness the extensive knowledge embedded in resource-rich CNNs, we introduce a novel heterogeneous teacher-student knowledge distillation framework based on LT-SpikingFormer. We select the classic ResNet-32 as the teacher model and train it with weak data augmentation. This selection is guided by the unique computational attributes of LT-SpikingFormer, enabling more flexible and efficient adaptation and optimization of the knowledge distillation process. The student model aims to achieve enhanced performance while maintaining lower complexity and fewer parameters.

For the heterogeneous architecture distillation between CNNs and SNNs, we emphasize extracting more profound and more effective information from the teacher model to improve the spiking neural network. We utilize two primary distillation strategies from global to local perspectives: Logits-based Global Distillation and Norm-reduced Local Distillation.

Logits-based Global Distillation. In the knowledge distillation training framework, logits function as the essential knowledge representation, encapsulating the prediction distribution of the teacher model. The Logits-based Global Distillation aims to align the prediction of the teacher and student models. The output distribution of the teacher model across different classes guides the student model's predictions, allowing the student to approximate the teacher's performance in the output space progressively. This approach is formalized as:

$$\mathcal{L}_{glo} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{D}_{KL}(y_n^t \| y_n^s), \quad (10)$$

where $\mathcal{D}_{KL}$ is the Kullback-Leibler divergence function, N is the total number of samples, y_n^s and y_n^t represent the outputs of the student and teacher models on the n-th sample, respectively. This approach enhances the student network's predictive accuracy, particularly for minority class samples, as the teacher model provides a more refined and comprehensive category distribution.

Norm-guided Local Distillation. To leverage the deep feature information from the teacher model more effectively, we aim to guide the student model in learning the feature representations of the teacher model across different network layers via multi-granularity distillation. However, due to the heterogeneous nature of the teacher and student models, their feature representations exhibit differences. Specifically, the teacher network produces dense, continuous-valued features, while the student network generates sparse, pulsed features. Direct alignment of these disparate features using convolutional operations could introduce superfluous information, leading the student model to learn irrelevant or secondary features and failing to capture the essential features of the teacher model. This misalignment may negatively impact the efficiency and generalization performance of the model.

To address this challenge, we propose a Norm-reduced Local Distillation module. The core idea is to project heterogeneous features from the student and teacher networks into a shared latent space using norm constraints to facilitate the alignment of heterogeneous features. We first extract the mean feature of each category from the teacher network, which can eliminate the impact of long-tailed data on feature distribution. For any sample x_n, the mean feature $p(x_n)$ of its category can be expressed as:

$$p(x_n) = \frac{1}{|C(x_n)|}\sum_{i=1}^{|C(x_n)|} \mathcal{T}(x_i), x_i \in X \& C(x_i) = C(x_n), \quad (11)$$

where $\mathcal{T}(\cdot)$ denotes the final output of the teacher network.

$C(x_n)$ indicates the category to which x_n belongs. Based on the above mean features, we use norm reduction to achieve mapping alignment of heterogeneous features. The implementation is defined as follows:

$$\mathcal{L}_t = \frac{1}{N}\sum_{n=1}^{N}\left|\left(\overline{\Theta(p(x_n))}+\Delta\Theta\right)-\|\Theta(\mathcal{T}(x_n))\|\right|, \quad (12)$$

$$\mathcal{L}_s^l = \frac{1}{N}\sum_{n=1}^{N}\left|\left(\overline{\Theta(p(x_n))}+\Delta\Theta\right)-\|\Phi(\mathcal{F}^l(x_n))\|\right|, \quad (13)$$

where $\mathcal{F}^l(\cdot)$ represents the feature output of the l-th latent layer of the student network, and $\Delta\Theta$ is the adjustment term for enhancement. Since the teacher network is pre-trained, its feature distribution is known and relatively stable, allowing us to set a benchmark value $\overline{\Theta(p(x_n))}$. The norm constraint minimizes the difference between each sample's mapped feature and the benchmark value, ensuring that the teacher's feature remains stable after being mapped to the latent space by the functions Θ and Φ.

For student loss $\mathcal{L}_s^l$, the primary objective of the norm constraint is to make the intermediate layer output of the student network closely approximate the mapped features of the teacher network. This is achieved by minimizing the difference between the norm of the student features after mapping and the teacher benchmark value. By jointly optimizing these loss functions, the student network is trained to project its features into the same latent space as the teacher network, thereby aligning heterogeneous feature representations. Of course, we can capture different latent features to achieve a heterogeneous alignment. The combined local distillation loss $\mathcal{L}_{loc}$ is expressed as:

$$\mathcal{L}_{loc} = \mathcal{L}_t + \frac{1}{L}\sum_{l=1}^{L}\mathcal{L}_s^l, \quad (14)$$

where L indicates the number of selected latent layer features. While local distillation shares some similarities with feature-based knowledge distillation, the norm-guided distillation in our LT-SpikingFormer is unique. Traditional feature-based distillation mainly aims to minimize the distance between intermediate feature layers. In contrast, our norm-guided local distillation uses norm constraints to project heterogeneous features into a shared latent space. This is new considering the differences between SNNs and CNNs. As equations show, it lessens the impact of long-tailed data on feature distribution and aligns student and teacher network features well.

Finally, the backbone network is trained using cross-entropy loss, which is defined as follows:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\bar{y}_{ij}\log(y_{ij}), \quad (15)$$

where N and C represent the number of samples and the number of classes, respectively. $\bar{y}_{ij}$ is the actual label for the i-th sample belonging to the j-th class. y_{ij} is the prediction made by the student model that the i-th sample belongs to the j-th class.

Overall, the above three loss functions during the representation learning stage are assembled as a whole optimization objective:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{glo}\mathcal{L}_{glo} + \lambda_{loc}\mathcal{L}_{loc}, \quad (16)$$

where λ_{glo} and λ_{loc} are the adjustable loss weight coefficients. This unified loss function integrates the contributions from both distillation strategies, promoting the effective training of the LT-SpikingFormer in the context of long-tailed data challenges.

4 Experiments

4.1 Experimental Setup

In this section, we conduct comprehensive experiments to evaluate the proposed LT-SpikingFormer framework and validate its efficacy in the context of long-tailed learning tasks. Initially, a series of analytical experiments were undertaken to substantiate our hypotheses and provide an in-depth examination of the framework's constituent components. Specifically, these analyses encompassed: 1) an overall performance evaluation of the model, 2) a comparative analysis of two distinct distillation strategies, 3) a sensitivity analysis of hyperparameters, and 4) an assessment of the impact of data augmentation strategies. Subsequently, we performed a thorough comparison of our proposed method with prevailing supervised learning approaches on benchmark long-tailed datasets, including CIFAR10/100-LT and ImageNet-LT.

Model. We propose the LT-SpikingFormer framework, which utilizes SpikingFormer, a spiking neural network based on the self-attention mechanism, as the backbone for the student model. For small-scale datasets like CIFAR10-LT and CIFAR100-LT, we use ResNet-32 as the teacher model, and for larger datasets like ImageNet-LT, we employ ResNet-50. During the training of the student model, we apply cross-entropy (CE) loss for the independent classifier and weighted Kullback-Leibler (KL) divergence for the distillation classifier to optimize the knowledge distillation process.

To further improve the student model, we propose a novel feature mapping strategy that uses norm constraints. This strategy maps the output of the teacher model and the intermediate-level information from the student model into a shared latent space. The student model, based on the Spike-Transformer framework, is divided into three equal parts, and the output of each segment is used as the exit point for feature mapping. Each exit point is made up of spike blocks, which are essential for transferring information from the teacher to the student.

Datasets. For the datasets, we assess our method using two small-scale datasets, CIFAR10-LT and CIFAR100-LT, and one large-scale dataset, ImageNet-LT. Long-tailed versions of CIFAR-10 and CIFAR-100 were generated using the approach described in [Cao *et al.*, 2019], where the imbalance factor β is the ratio of the sample size of the most frequent class to the least frequent class. In our experiments, β was set to 100, 50, and 10, respectively. The original validation sets of CIFAR10 and CIFAR100 were used directly for testing. For ImageNet, we generated the long-tailed variant by sampling according to a Pareto distribution with a power value α_p

Method	Architecture	CIFAR10-LT			CIFAR100-LT		
		$\beta = 100$	$\beta = 50$	$\beta = 10$	$\beta = 100$	$\beta = 50$	$\beta = 10$
Single-Stage Training							
CB-Focal [Cui *et al.*, 2019]	ResNet-32	74.57	79.27	87.49	39.60	45.32	57.99
LDAM-DRW [Cao *et al.*, 2019]	ResNet-32	77.03	79.32	88.16	42.04	45.11	58.71
LDAM-DAP [Jamal *et al.*, 2020]	ResNet-32	80.0	82.3	87.4	44.08	49.16	58.00
LA [Menon *et al.*, 2020]	ResNet-32	77.7	-	-	43.9	-	-
IBLLoss [Park *et al.*, 2021]	ResNet-32	77.97	82.38	87.90	44.96	48.92	59.54
MiSLAS [Rangwani *et al.*, 2022]	ResNet-32	82.1	85.7	90.0	47.0	52.3	63.2
VS+SAM [Wei *et al.*, 2023]	ResNet-32	82.4	-	-	46.6	-	-
LCReg [Liu *et al.*, 2024]	ResNet-32	83.1	86.5	91.2	47.6	53.1	64.2
ProCo [Du *et al.*, 2024]	ResNet-32	85.9	88.2	91.91	52.8	57.1	65.5
Multi-Stage Training							
MDCS [Zhao *et al.*, 2023]	ResNet-32	85.8	89.4	-	46.0	50.5	62.3
BKD [Zhang *et al.*, 2023]	ResNet-32	81.72	83.81	89.21	45.00	49.64	61.33
NCL++ [Tan *et al.*, 2024]	ResNet-32	86.1	88.0	-	54.8	58.2	-
SpikingFormer Backbone							
SpikingFormer	SpikingFormer	72.04	77.51	85.69	39.08	41.77	46.00
SpikingFormer+Focal [Lin, 2017]	SpikingFormer	74.11	81.18	90.01	41.74	47.96	49.03
SpikingFormer+DRW	SpikingFormer	76.74	82.52	88.24	43.68	48.73	52.09
LT-SpikingFormer (ours)	SpikingFormer	**86.71**	**89.10**	**93.97**	**55.94**	**60.66**	**67.53**

Table 1: Results on CIFAR10-LT and CIFAR100-LT datasets with β=10, β=50 and β=100.

= 6. The classes were categorized into Head, Medium, and Tail groups, and classification results were computed separately for each group.

For data augmentation, CIFAR10/100-LT datasets used weak augmentation techniques such as cropping, horizontal flipping, and rotation. Strong augmentation included these basic methods along with CIFAR10Policy and mixup [Zhang, 2017]. For ImageNet-LT, weak augmentation consisted of cropping, horizontal flipping, rotation, and ColorJitter. Strong augmentation for ImageNet-LT included mixup and cutmix [Yun *et al.*, 2019] to ensure a fair comparison.

Baselines. In our experiments, we established multiple baselines to assess the performance of LT-SpikingFormer. The primary baseline was the SpikingFormer network itself, which served as the backbone of the student model across all experiments. This allowed for a consistent comparison of improvements introduced by LT-SpikingFormer. Additionally, we integrate the backbone model with state-of-the-art methods designed for long-tailed data distributions and compare their performance with that of our proposed method.

Optimization. For optimization, different optimizers were used for the student and teacher models. The student model employed AdamW with a weight decay of 0.01, a batch size of 128, and a cosine annealing learning rate starting at 0.0005 for 600 epochs on the CIFAR10/100-LT dataset. For ImageNet-1K, the learning rate was 0.001, and training lasted 320 epochs. The teacher model, based on CNN, was optimized using Sharpness-Aware Minimization (SAM) to improve generalization.

4.2 Result

We evaluate the proposed LT-SpikingFormer method against both single-stage and multi-stage training approaches using the CIFAR10-LT, CIFAR100-LT, and ImageNet-LT datasets.

Method	Architecture	ImageNet-LT			
		Head	Medium	Tail	All
Single-Stage Training					
CB-Focal [Cui *et al.*, 2019]	ResNet-50	39.6	32.7	16.8	33.2
cRT [Kang *et al.*, 2020]	ResNet-50	62.5	47.4	29.5	50.3
LDAM-DRW [Cao *et al.*, 2019]	ResNet-50	61.1	48.2	28.3	49.9
LA [Menon *et al.*, 2020]	ResNet-50	61.1	47.5	27.6	50.1
MiSLAS [Zhong *et al.*, 2021]	ResNet-50	62.9	50.7	34.3	52.7
LDAM+DRW+SAM [Rangwani *et al.*, 2022]	ResNet-50	62.0	52.1	34.8	53.1
RBL [Peifeng *et al.*, 2023]	ResNet-50	64.8	49.6	34.2	53.3
LCReg [Liu *et al.*, 2024]	ResNet-50	-	-	-	55.3
ProCo [Du *et al.*, 2024]	ResNet-50	68.2	55.1	38.1	57.8
Multi-Stage Training					
DiVE [He *et al.*, 2021]	ResNeXt-50	64.06	50.41	31.46	53.10
BKD [Zhang *et al.*, 2023]	ResNet-152	54.6	37.2	20.4	41.6
NCL++ [Tan *et al.*, 2024]	ResNet-50	-	-	-	58.0
SpikingFormer Backbone					
SpikingFormer	Spikingformer	61.31	51.08	30.36	51.14
SpikingFormer+Focal[Lin, 2017]	SpikingFormer	60.92	52.20	35.84	52.38
SpikingFormer+DRW	SpikingFormer	62.39	54.00	41.81	54.72
LT-SpikingFormer(ours)	SpikingFormer	**65.40**	**59.68**	**43.59**	**58.66**

Table 2: The overall Top-1 accuracy, as well as the Top-1 accuracy for the Head, Medium and Tail, on the ImageNet-LT dataset.

All the methods examined are specifically designed to handle long-tailed data distributions.

Training results on small-scale. The table 1 present the training outcomes for CIFAR10-LT and CIFAR100-LT with imbalance factors of 10, 50, and 100, respectively. The proposed LT-SpikingFormer method achieves substantial performance improvements over competing approaches. Specifically, the LT-SpikingFormer model outperforms the backbone model by 14.67% on CIFAR10-LT and 16.86% on CIFAR100-LT. When compared to the distillation method BKD, our results show a 5%-10% improvement on both datasets. Notably, the common multi-stage training strategy in long-tailed learning typically involves the use of expert models. Our method demonstrates approximately 1% higher performance than MDCS, which is based on a self-distillation

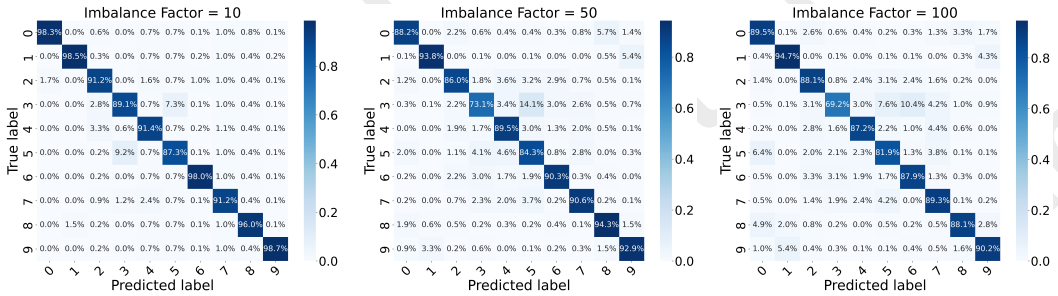


Figure 4: Confusion matrices for our method applied to the CIFAR10-LT dataset with imbalance factors of 10, 50 and 100.

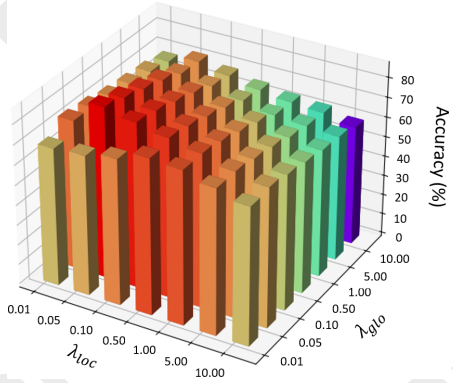


Figure 5: Experiments on the influence of loss weight λ_{glo} and λ_{loc} distribution on model performance.

w/MixUp	$\mathcal{L}_{ce}$	$\mathcal{L}_{glo}$	$\mathcal{L}_{loc}$	Accuracy(%)
✗	✓	✗	✗	72.04
✓	✓	✗	✗	76.54
✓	✓	✓	✗	80.99
✓	✓	✗	✓	83.62
✓	✓	✓	✓	86.71

Table 3: An ablation study was conducted on the CIFAR10-LT dataset with an imbalance factor (β) of 100. The term w/MixUp refers to the use of weak-strong augmentation. The cross-entropy loss ($\mathcal{L}_{ce}$) serves as a fundamental component of the network and is therefore not used for ablation comparison experiments. $\mathcal{L}_{glo}$ represents the Logits-based Global loss, and $\mathcal{L}_{loc}$ corresponds to the Norm-reduced Local loss.

expert model, on CIFAR10-LT, with similar improvements observed on CIFAR100-LT.

Training results on large-scale. In table 2, we present the overall Top-1 accuracy for ImageNet-LT, along with the Top-1 accuracy for head, medium and tail categories. The LT-SpikingFormer outperforms other competing methods on this dataset. To ensure a fair comparison, we re-implemented certain multi-stage networks under a consistent training environment. On the ImageNet-LT dataset, our LT-SpikingFormer model demonstrates a 7.52% improvement over the backbone model. Compared to existing methods such as DiVE and BKD, our LT-SpikingFormer achieves state-of-the-art (SOTA) performance under identical conditions.

4.3 Ablation Study and Further Analysis

Ablation studies on all components. As detailed in table 2, we evaluate the proposed components, including the Logits-based Global loss ($\mathcal{L}_{glo}$), Norm-reduced Local loss ($\mathcal{L}_{loc}$), and weak-strong augmentation (w/mixup). When both $\mathcal{L}_{glo}$ and $\mathcal{L}_{loc}$ are marked as ✗ and $\mathcal{L}_{ce}$ is ✓, it indicates that only the backbone model, SpikingFormer, is used for training. The ✗ in w/MixUp signifies that the data augmentation applied is limited to weak augmentation. As shown in table 3, implementing the LT-SpikingFormer with a combination of strong and weak augmentations increases performance from 72.04% to 76.54%, demonstrating the efficacy of mixup in handling long-tailed distributions. When the teacher model is introduced and the $\mathcal{L}_{glo}$ and $\mathcal{L}_{loc}$ losses are applied during training, the overall model performance improves by 4.45% and 7.08%, respectively. Lastly, applying the full distillation loss results in a further significant performance boost, raising accuracy to 86.71%.

Overview of the Imbalanced Classification Performance. Fig. 4 presents three confusion matrices for our method applied to the CIFAR-LT dataset, with imbalance factors of 100, 50, and 10, respectively. As depicted in the figure, our method performs effectively, particularly when the imbalance factor is large. As the imbalance factor increases, the method maintains high classification accuracy, with prominent values along the diagonal indicating robust prediction accuracy across all classes. Even at an imbalance factor of 100, the misclassification rate remains low, and the off-diagonal values are minimal. These results demonstrate that our method achieves excellent performance even in the presence of highly imbalanced datasets, highlighting its strong robustness and applicability.

Influence of loss weight λ_{glo} and λ_{loc}. λ_{glo} and λ_{loc} represent the weights of the two loss functions, and are used to control their relative contributions to the total loss. To determine suitable values for λ_{glo} and λ_{loc}, we performed a series of experiments on the CIFAR10-LT dataset (β=100). As illustrated in Fig. 5, the optimal performance was achieved when $\lambda_{glo} = 0.5$ and $\lambda_{loc} = 0.05$. This result indicates that our model has effectively balanced the use of external guidance for knowledge transfer, the refinement of internal feature representations, and its ability to learn directly from the data, thereby achieving the best overall performance.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China (No.2023YFC3305600), Joint Fund of Ministry of Education of China (8091B02072404), National Natural Science Foundation of China (62132016, 62171343, and 62201436), Key Research and Development Program of Shaanxi (2024GX-YBXM-127), Natural Science Basic Research Program of Shaanxi (2020JC-23) and National Key Laboratory Foundation of China (Grant No. HTKJ2024KL504011).

References

[Ahn *et al.*, 2019] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171, 2019.

[Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[Chen *et al.*, 2024] Xiangru Chen, Chenjing Liu, Peng Hu, Jie Lin, Yunhong Gong, Yingke Chen, Dezhong Peng, and Xue Geng. Adaptive masked autoencoder transformer for image classification. *Applied Soft Computing*, 164:111958, 2024.

[Cho and Hariharan, 2019] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[Diehl *et al.*, 2015] 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.

[Du *et al.*, 2024] Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Hao *et al.*, 2023] Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 79570–79582. Curran Associates, Inc., 2023.

[He and Garcia, 2009] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[He *et al.*, 2021] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 235–244, 2021.

[Jamal *et al.*, 2020] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7610–7619, 2020.

[Jin *et al.*, 2023] Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23695–23704, 2023.

[Kang *et al.*, 2020] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020.

[Khan *et al.*, 2017] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[Li *et al.*, 2022] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6918–6928, 2022.

[Liang *et al.*, 2024] Ke Liang, Lingyuan Meng, Yue Liu, Meng Liu, Wei Wei, Suyuan Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. Simple yet effective: Structure guided pre-trained transformer for multimodal knowledge graph reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1554–1563, 2024.

[Lin, 2017] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[Liu *et al.*, 2024] Weide Liu, Zhonghua Wu, Yiming Wang, Henghui Ding, Fayao Liu, Jie Lin, and Guosheng Lin. Lcreg: Long-tailed image classification with latent categories based recognition. *Pattern Recognition*, 145:109971, 2024.

[Long *et al.*, 2022] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6959–6969, 2022.

[Ma *et al.*, 2024] Huimin Ma, Siwei Wang, Shengju Yu, Suyuan Liu, Jun-Jie Huang, Huijun Wu, Xinwang Liu, and En Zhu. Automatic and aligned anchor learning strategy for multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5045–5054, 2024.

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

[Menon *et al.*, 2020] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[Mostafa, 2017] Hesham Mostafa. Supervised learning based on temporal coding in spiking neural networks. *IEEE transactions on neural networks and learning systems*, 29(7):3227–3235, 2017.

[Neftci *et al.*, 2019] 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.

[Park *et al.*, 2021] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 735–744, 2021.

[Peifeng *et al.*, 2023] Gao Peifeng, Qianqian Xu, Peisong Wen, Zhiyong Yang, Huiyang Shao, and Qingming Huang. Feature directions matter: Long-tailed learning via rotated balanced representation. In *International Conference on Machine Learning*, pages 27542–27563. PMLR, 2023.

[Rangwani *et al.*, 2022] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, et al. Escaping saddle points for effective generalization on class-imbalanced data. *Advances in Neural Information Processing Systems*, 35:22791–22805, 2022.

[Sengupta *et al.*, 2019] 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.

[Shen *et al.*, 2025a] Jiangrong Shen, Kejun Wang, Wei Gao, Jian K Liu, Qi Xu, Gang Pan, Xiaodong Chen, and Huajin Tang. Temporal spiking generative adversarial networks for heading direction decoding. *Neural Networks*, 184:106975, 2025.

[Shen *et al.*, 2025b] Jiangrong Shen, Qi Xu, Gang Pan, and Badong Chen. Improving the sparse structure learning of spiking neural networks from the view of compression efficiency. *arXiv preprint arXiv:2502.13572*, 2025.

[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 (CVPR)*, 2024.

[Tan *et al.*, 2024] Zichang Tan, Jun Li, Jinhao Du, Jun Wan, Zhen Lei, and Guodong Guo. Ncl++: Nested collaborative learning for long-tailed visual recognition. *Pattern Recognition*, 147:110064, 2024.

[Tung and Mori, 2019] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019.

[Wang *et al.*, 2024] Boyue Wang, Guangchao Wu, Xiaoyan Li, Junbin Gao, Yongli Hu, and Baocai Yin. Modality perception learning-based determinative factor discovery for multimodal fake news detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Wei *et al.*, 2023] Xiu-Shen Wei, Xuhao Sun, Yang Shen, Anqi Xu, Peng Wang, and Faen Zhang. Delving deep into simplicity bias for long-tailed image recognition. *arXiv preprint arXiv:2302.03264*, 2023.

[Yao *et al.*, 2024] 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.

[Yun *et al.*, 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[Zhang *et al.*, 2023] Shaoyu Zhang, Chen Chen, Xiyuan Hu, and Silong Peng. Balanced knowledge distillation for long-tailed learning. *Neurocomputing*, 527:36–46, 2023.

[Zhang, 2017] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[Zhao *et al.*, 2023] Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. Mdcs: More diverse experts with consistency self-distillation for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11597–11608, 2023.

[Zhao *et al.*, 2024] Han Zhao, Xu Yang, Cheng Deng, and Junchi Yan. Dynamic reactive spiking graph neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16970–16978, 2024.

[Zhong *et al.*, 2021] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021.

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