Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Exploiting Label Skewness for Spiking Neural Networks in Federated Learning

Di Yu[1], **Xin Du**[1*], **Linshan Jiang**[2], **Huijing Zhang**[1] and **Shuiguang Deng**[1]
[1]Zhejiang University
[2]National University of Singapore
{yudi2023,xindu}@zju.edu.cn, linshan@nus.edu.sg, {huijingzhang, dengsg}@zju.edu.cn

Abstract

The energy efficiency of deep spiking neural networks (SNNs) aligns with the constraints of resource-limited edge devices, positioning SNNs as a promising foundation for intelligent applications leveraging the extensive data collected by these devices. To address data privacy concerns when deploying SNNs on edge devices, federated learning (FL) facilitates collaborative model training by leveraging data distributed across edge devices without transmitting local data to a central server. However, existing FL approaches struggle with label-skewed data across devices, which leads to drift in local SNN models and degrades the performance of the global SNN model. In this paper, we propose a novel framework called FedLEC, which incorporates intra-client label weight calibration to balance the learning intensity across local labels and inter-client knowledge distillation to mitigate local SNN model bias caused by label absence. Extensive experiments with *three* different structured SNNs across *five* datasets (i.e., *three* non-neuromorphic and *two* neuromorphic datasets) demonstrate the efficiency of FedLEC. Compared to *eight* state-of-the-art FL algorithms, FedLEC achieves an average accuracy improvement of approximately 11.59% for the global SNN model under various label skew distribution settings.

1 Introduction

Recent research has demonstrated that deep spiking neural networks (SNNs) are well-suited for deployment on resource-constrained edge devices [Di *et al.*, 2024], thanks to their energy efficiency [Maass, 1997] and compatibility with the computational limitations of these devices. These edge devices can also collect vast data, including photos, videos, and audio, which presents numerous opportunities with deep SNNs for influential research and practical applications [Lim *et al.*, 2020]. However, the traditional approach of centralizing this data has become progressively impractical, as it often involves sensitive information unsuitable for sharing, and the communication costs of transmitting large volumes of data are prohibitively high.

Federated Learning (FL) has emerged as a viable solution, enabling the development of deep SNN models while keeping data localized on edge devices [Venkatesha *et al.*, 2021]. In FL, edge devices, commonly called *clients*, train *local models* on their respective local datasets. These locally trained models are subsequently transmitted to the central FL *server*, aggregating them into a *global model* to facilitate real-world applications. However, the characteristics of the data distribution across edge devices can present significant challenges to federated SNN learning, including:

Label Skewness. This type of data heterogeneity is common in practice as edge devices may only possess a subset of the global categories in their acquired datasets according to their deployed environments. For example, the distribution of animal species varies across different geographical regions, such as pandas being native to China and koalas exclusive to Australia. When the corresponding label distributions across clients are extremely **imbalanced**, local models struggle to generalize to the global distribution, inducing a sub-optimal global model [Xia *et al.*, 2024].

Bias Aggravation. Although back-propagation with surrogate functions [Wu *et al.*, 2019] allows local SNN models to be efficiently trained by edge GPUs, it also introduces layer-wise accumulated gradient drifts [Deng *et al.*, 2023]. When SNNs are trained on highly skewed data, the absence of relevant knowledge about **missing labels** aggravates training drift, leading to severe gradient bias and increased local over-fitting. Hence, aggregating biased gradients from clients causes the global SNN model to converge poorly, ultimately degrading its performance [Wang *et al.*, 2023].

To mitigate these issues, we propose a novel FL framework named FedLEC[1]. ***First***, FedLEC introduces a calibrated objective function based on the corresponding probability of local label occurrences. By directing the training process to highlight the margins of majority and minority labels to achieve optimal thresholds, FedLEC promotes larger margin attainment for underrepresented minority labels, thereby enhancing their representation within the model and balancing the intra-client inconsistency. ***Second***, we

*Corresponding Authors: Xin Du and Shuiguang Deng.

[1]https://github.com/AmazingDD/FedLEC

employ a label distillation method to convey global distribution information regarding missing labels across clients. This approach ensures the local outputs align closely with the global model's predictions for missing labels, effectively mitigating the aggravated bias and over-fitting issues arising from label absence. Besides, FedLEC focuses on alleviating the impact of locally skewed data so that it can be integrated with other methods [Karimireddy *et al.*, 2020; Li *et al.*, 2020] that address inter-client objective inconsistency to improve the performance of the server model further.

To validate the effectiveness of our FedLEC, we conduct extensive experiments under various label skew settings, substantiating that FedLEC significantly improves the accuracy of federated SNN learning compared to other state-of-the-art (SOTA) FL algorithms under label skews, particularly under extreme label skew conditions. The primary contributions of this paper are summarized as follows:

- We first analyze the impact of label skewness in federated SNN learning and demonstrate that applying FL methods designed for artificial neural networks (ANNs) under extreme label skew conditions is suboptimal. This analysis lays the foundation for developing specialized solutions tailored to SNNs.
- We propose FedLEC, a federated learning framework designed to address the challenges posed by extreme label skew. FedLEC enhances the generalization ability of local SNN models by balancing intra-client label learning and mitigating bias from missing labels through inter-client knowledge alignment. This improves the global model's overall performance.
- Results from extensive experiments with **three** differently structured SNNs and **eight** FL baselines across **five** datasets demonstrate the efficacy of FedLEC. Under different extreme label skews, FedLEC can outperform the other baselines on average accuracy by about 11.59% across all empirical trials in this study.

2 Related Work

Researchers have investigated how to leverage the energy efficiency of SNNs in FL systems [Venkatesha *et al.*, 2021]. Previous studies mainly focused on federated SNN learning across various application scenarios. For instance, FedSNN-NRFE [Xie *et al.*, 2022] exploited privacy-preserving FL, training SNN models for traffic sign recognition on the Internet of Vehicles. A distributed FL system with SNNs [Zhang *et al.*, 2024] was built to process radar data collaboratively. SURFS [Aouedi and Kandaraj, 2024] proposed a robust and sustainable instruction detection system with federated SNN learning. Some studies also explore the compatibility of different FL architectures with SNNs. Hierarchical FL [Aouedi *et al.*, 2023; Aouedi and Kandaraj, 2024] was utilized with SNNs to lower the communication latency and enhance model robustness.

However, these studies overemphasize the application implementation by combining FL with SNNs and neglect the inherently heterogeneous data risks in real-world systems. Although several studies [Venkatesha *et al.*, 2021; Tumpa *et al.*, 2023] have preliminarily explored the impact of non-IID data on federated SNN learning, the influence of one common non-IID data heterogeneous challenge called label skew is often neglected by most researchers. Meanwhile, few studies have investigated whether solutions for label skews in FL applied to ANNs like FedLC [Zhang *et al.*, 2022] and FedConcat [Diao *et al.*, 2024] remain the same effectiveness for SNNs. Motivated by current research limitations, this paper delves into tackling the significant drop in accuracy performance observed when implementing federated SNN learning under label skew.

3 Preliminary

3.1 Federated Learning

An FL system comprises M local client nodes and a central global server. The server initiates the federated training process by broadcasting the initial model parameters θ to selected client nodes. Subsequently, the locally trained model at each client m is updated using the client's private data shard D^m. The model update from each client comprises the accumulated gradients throughout local training. These updates are periodically communicated to the server for aggregation. In one communication round, the updated parameters θ^m from client m are transmitted to the server for global model updating. The server will then aggregate these perceived parameters in a specific manner to update the global model. A standard aggregation method using the weighted average is FedAvg [McMahan *et al.*, 2017].

3.2 Spiking Neural Network

Leaky Integrate-and-Fire (LIF) neuron, widely used in diverse domains [Yao *et al.*, 2024; Su *et al.*, 2023], is one of the popular spiking neurons utilized to build spiking neural networks, whose computational paradigm is:

$$V[t^-] = V[t-1] + \beta(I[t] + V[t-1]) \tag{1}$$

$$S[t] = H(V[t^-] - \overline{V}) \tag{2}$$

$$V[t] = S[t]V_r + (1 - S[t])V[t^-] \tag{3}$$

where $\beta, \overline{V}$, and V_r represent the membrane time constant, the firing threshold, and the reset membrane potential, respectively. At time step t, $I[t]$ is the spatial input, $V[t^-]$ and $V[t]$ are the membrane potential after neuronal dynamics integration and after the trigger of firing separately. $H(\cdot)$ is the Heaviside step function, generating a binary spike when $V[t^-] \geq \overline{V}$. Subsequently, $V[t]$ will reset to V_r or remain unchanged otherwise. To tackle the non-differentiable issue of Equation (2), studies [Wu *et al.*, 2019; Deng *et al.*, 2023] introduce back-propagation through time (BPTT) with surrogate functions to train SNNs efficiently on GPUs.

4 Methodology

4.1 Problem Statement

There are two different label skew settings in FL [Li *et al.*, 2022]: *quantity-based* and *distribution-based*, as depicted in Figure 1. For *quantity-based* label skew, each client can only

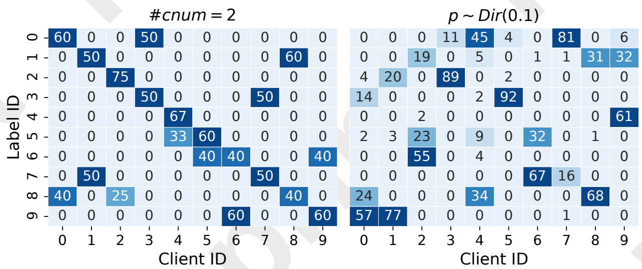

Figure 1: Different types of label skews. Cell number represents the percentage value of the samples of a specific class allocated.

Figure 2: The impact of label skewness on local updating for models in three selected clients. The histogram indicates the number of samples for each label in the local data shard.

hold samples of fixed c different labels. Samples with a specific label will be randomly divided into equal data shards and distributed to clients who own this label so that no overlap exists among the data shards of different clients (for simplicity, we use $\#cnum = c$ to denote this skew). The other type of label imbalance is *distribution-based label skew* [Yurochkin *et al.*, 2019], which allocates portions of the samples for each label according to Dirichlet distribution [Huang, 2005]. In particular, this skew allocates $p_{c,m}\%$ samples of class c to client m, where $p_c \sim Dir(\alpha)$ and α (> 0) is the concentration parameter to control the imbalance level. In this study, we utilize $p \sim Dir(\alpha)$ to represent this type of label skew.

Definition 1 (Majority, Minority and Missing Labels)**.** *For a label-skewed local data shard D, its label set $\mathcal{C}$ can be divided into three categories: the majority label set $\mathcal{J}$, the minority label set $\mathcal{K}$, and the missing label set $\mathcal{M}$. We have $\mathcal{C} = \mathcal{J} \cup \mathcal{K} \cup \mathcal{M}$, $\mathcal{J} \cap \mathcal{K} = \emptyset$, $\mathcal{J} \cap \mathcal{M} = \emptyset$, $\mathcal{M} \cap \mathcal{K} = \emptyset$, where $\emptyset$ is an empty set. The number of samples $(\mathbf{x}, y) \sim D$ with label set $\mathcal{C}$ satisfies $|D_{\mathcal{J}}| \gg |D_{\mathcal{K}}| > |D_{\mathcal{M}}| = 0$.*

In FL, a total of M clients aim to minimize the below objective with model parameter θ:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \sum_{m=1}^{M} \mathcal{P}_m \cdot \mathcal{L}_m(\theta) \tag{4}$$

where $\mathcal{L}_m$ is the objective function for local model in the m-th client, $\mathcal{P}_m$ is the relative local sample size, and $\sum_{m=1}^{M} \mathcal{P}_m = 1$. The local data distribution P_m varies across clients for label skew settings, resulting in significantly different models after local training.

Besides, we empirically validate the impact of label skew in federated SNN learning on three selected clients with *cifar10* dataset after sufficient communication rounds. After verifying that the current local models are consistent with the global SNN model, we evaluate the accuracy for each label both before and after the local update. As shown in Figure 2, after local training is conducted separately, the test accuracy for majority labels typically improves compared to the initial state. However, the test accuracy for minority and missing labels significantly decreases, often approaching zero. This observation highlights that label skewness can result in a biased SNN model .

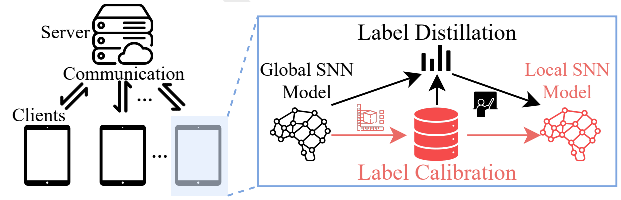

Figure 3: The framework of FedLEC.

4.2 Framework Overview

We propose a novel **fed**erated SNN learning framework with **l**abel sk**e**w **c**alibration (FedLEC). As shown in Figure 3, we conduct two strategies to alleviate the intra-client label inconsistency. ***First***, we calibrate the logits[2] of each label before applying the softmax cross-entropy by adjusting them based on the occurrence probability of each label, thereby mitigating the impact of intra-client data imbalance from minority labels during the local update. ***Second***, the global model from the previous communication round, serving as a teacher model, transfers knowledge about the inter-client data distribution to each local model, thereby mitigating over-fitting issues related to missing labels. Then, each local objective function $\mathcal{L}_m$ in Equation (4) for one time step is:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}_m(\theta) = (1 - \lambda) \cdot \mathcal{L}_c(\theta) + \lambda \mathcal{L}_d(\theta) \tag{5}$$

where $\mathcal{L}_c$ and $\mathcal{L}_d$ represent the objective functions derived from the two strategies, respectively. By adjusting the hyper-parameter λ, the model balances its focus between addressing intra-client and inter-client bias during local learning.

Label Calibration. Denote the data distribution at m-th client as $P(\mathbf{x}, y) = P(\mathbf{x}|y)P(y)$. For a sample $\mathbf{x}$, the predicted label is $\hat{y} = \arg\max_y f(\mathbf{x}; \theta)$, where $f(\mathbf{x}; \theta)$ is the corresponding logits from the local model. In the one communication round of the FL process, the server broadcasts the global model parameters to the selected local client m. The goal of the local training task is to minimize the classification error from a statistical perspective:

$$\min P(y \neq \hat{y}|\mathbf{x}) \tag{6}$$

According to Bayes rule, Equation (6) is proportional to $P(\mathbf{x}|y)P(y)$. When the label distribution is balanced, i.e., $P(y)$ is equivalent for all label y, then the soft-max cross-entropy can be regarded as the surrogate loss function for Equation (6) since the probability $P(\mathbf{x}|y) \propto e^{f(\mathbf{x};\theta)}$. However, when the local data suffer from label distribution skewness, $P(y)$ is no longer identical, i.e., minority labels have

[2]Logits refer to the output of the model's final classification layer

much lower probabilities of occurrence than majority labels, inspiring us to calibrate the error in Equation (6) via averaging the per-label error rate [Menon *et al.*, 2021] like: $\frac{1}{|\mathcal{C}|}\sum_{y\in\mathcal{C}} P(\mathbf{x}|y \neq \hat{y})$. In this manner, varying $P(y)$ due to label skewness will no longer affect the results. To minimize the 'balanced' calibrated error, we, in turn, need to find:

$$\begin{aligned}&\arg\max_{y\in\mathcal{C}} P(\mathbf{x}|y=\hat{y}) = \arg\max_{y\in\mathcal{C}} P(y|x)/P(y)\\ &= \arg\max_{y\in\mathcal{C}}(f_y(\mathbf{x};\theta) - \log(\gamma_y))\end{aligned} \tag{7}$$

where γ_y is the estimation of the prior $P(y)$ for a specific label in the local label set $\mathcal{C}$. The logits of minority labels in $\mathcal{C}$ are reduced by a relatively larger value, whereas those of majority labels are adjusted oppositely, thereby balancing the classification loss. Inspired by Equation (7), we can then alleviate the impact of intra-client label inconsistency by training the local SNN model with a calibrated cross-entropy loss:

$$\mathcal{L}_c = -\mathbb{E}_{(\mathbf{x},y)\sim D}\log\left(\frac{e^{f_y(\mathbf{x};\theta)} - \gamma_y}{\sum_{c\in\mathcal{C}}(e^{f_c(\mathbf{x};\theta)} - \gamma_c)}\right) \tag{8}$$

Notably, $\mathcal{L}_c$ prioritize mitigating the biases between the majority and minority labels. However, it fails to handle missing label issues since no related class prior is available. Hence, it is necessary to design distinct objective functions to ensure that the global model retains its original predictive capability for missing labels, preventing the loss of such capability due to over-fitting caused by the absence of relevant data.

Label Distillation. Figure 1 has demonstrated that extreme quantity-based and distribution-based label skews often induce numerous labels missing in local data shard, i.e., $|\mathcal{M}| \gg |\mathcal{J}|+|\mathcal{K}|$. After implementing local training, each client's local model will be over-fitted according to the local data distribution. The core idea to alleviate this issue is to *preserve the global view of the data distribution for locally missing labels*. Specifically, when the server aggregates gradients from the local models during each communication round, the global could learn and retain distribution information about all the labels. Therefore, we employ the predictions of the global model on locally available data to perform knowledge distillation (KD) for local missing labels in $\mathcal{M}$. Besides, KD can effectively mitigate the over-fitting issue of SNNs caused by accumulated gradient errors [Zuo *et al.*, 2024]. Then, FedLEC defines the distillation loss $\mathcal{L}_d$ as the KL-divergence loss between the soft-max prediction vector of the global model $\tilde{q}_g$ and that of the local model $\tilde{q}_l$ as follows:

$$\mathcal{L}_d = \mathbb{E}_{(\mathbf{x},y)\sim D}\sum_{c\in\mathcal{M}} \tilde{q}_g[c]\log\frac{\tilde{q}_g[c]}{\tilde{q}_l[c]} \tag{9}$$

which guarantees the local model effectively aligns with the global model's predictive capabilities for missing labels.

Aggregation Through Time. According to Equation (4), the aggregate process of SNN model parameters in FedLEC at one communication round can be expressed as:

$$\theta = \underbrace{\sum_{m=1}^{M}\mathcal{P}_m\left(\frac{1}{T}\sum_{t=1}^{T}\theta^t\right)}_{\text{FedLEC}} = \frac{1}{T}\sum_{t=1}^{T}\underbrace{(\sum_{m=1}^{M}\mathcal{P}_m\theta^t)}_{\text{FedAvg}} \tag{10}$$

Algorithm 1 Aggregation Through Time in FedLEC

Input: Selected Client Set $\mathcal{S}$, Total Timesteps T
Output: Updated SNN Parameters θ.

Server executes:
for all $t = 1, ..., T$ **do**
  Receive parameters $\theta^{m,t}$ from each client m.
  $\theta^t \leftarrow \sum_{m\in\mathcal{S}}\frac{|D^m|}{\sum_{m\in\mathcal{S}}|D^m|}\theta^{m,t}$
end for
$\theta \leftarrow \frac{1}{T}\sum_{t=1}^{T}\theta^t$.
return θ

where T is the total time steps the SNN model training process requires. Intuitively, the global aggregation process of FedLEC scales with time steps compared to conventional FedAvg. We term this parameter aggregation process in federated SNN learning as Aggregation Through Time (ATT), whose workflow is elaborated in Algorithm 1.

4.3 Convergence Analysis

Inspired by [Li *et al.*, 2019], we make the following assumptions on the objectives $\mathcal{L}_1, ..., \mathcal{L}_M$. Assumptions 1 and 2 are standard; a typical example is the soft-max classifier. Since the KL-divergence loss $\mathcal{L}_d$ and the calibrated cross-entropy $\mathcal{L}_c$ are typical variants of the soft-max classifier, the local objective $\mathcal{L}_m$ of FedLEC also satisfies these assumptions.

Assumption 1 (Each $\mathcal{L}_m$ is L-smooth). *For all* $\mathbf{v}$ *and* $\mathbf{w}$, $\mathcal{L}_m(\mathbf{v}) \leq \mathcal{L}_m(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T\nabla\mathcal{L}_m(\mathbf{w}) + \frac{L}{2}||\mathbf{v}-\mathbf{w}||_2^2$

Assumption 2 (Each $\mathcal{L}_m$ is μ-strongly convex). *For all* $\mathbf{v}$ *and* $\mathbf{w}$, $\mathcal{L}_m(\mathbf{v}) \geq \mathcal{L}_m(\mathbf{w}) + (\mathbf{v}-\mathbf{w})^T\nabla\mathcal{L}_m(\mathbf{w}) + \frac{\mu}{2}||\mathbf{v}-\mathbf{w}||_2^2$

Let the number of local iterations performed in one device between two communications be E, and the total number of every device's SGDs be R. Given that Equation (10) ensures R is evenly divisible by the number of time steps T, we simplify the subsequent analysis by omitting T, as its effect is already encapsulated within R. Note that with surrogate functions [Wu *et al.*, 2019], SNNs can implement backpropagation with gradient descent like ANNs. Hence, for the r-th SGD of the local SNN model in client m, we have:

Assumption 3 (Variance boundary). *Let* ξ_r^m *be randomly sampled from the* m*-th client's local data. Each device's variance of stochastic gradients is bounded:* $\mathbb{E}||\nabla\mathcal{L}_m(\theta_r^m, \xi_r^m) - \nabla\mathcal{L}_m(\theta_r^m)||^2 \leq \sigma_m^2$, *for* $m = 1, ..., M$.

Assumption 4. *The expected squared norm of stochastic gradients is uniformly bounded, i.e.,* $\mathbb{E}||\nabla\mathcal{L}_m(\theta_r^m, \xi_r^m)||^2 \leq G^2$ *for all* $m = 1, .., M$ *and* $r = 1, ..., R-1$.

Let $\mathcal{L}^*$ and $\mathcal{L}_m^*$ be the minimum values of $\mathcal{L}$ and $\mathcal{L}_m$, respectively. We utilize the term $\Gamma = \mathcal{L}^* - \sum_{m=1}^{M}\mathcal{P}_m\mathcal{L}_m^*$ to quantify the degree of data heterogeneity (label skewness). If the data is IID, Γ approaches zero as the number of samples increases. If the data is non-IID, then Γ is nonzero, and its magnitude reflects the skewness of the data distribution.

We first analyze the case that all the clients participate in the parameter aggregation step. From Equation (10), we acknowledge that the aggregation step in FedLEC for each time step t is equivalent to that of FedAvg, denoted as:

$$\theta^m_{r+i+1} \leftarrow \theta^m_{r+i} - \eta_{r+i} \nabla \mathcal{L}_m(\theta^m_{r+i}, \xi^m_{r+i}) \tag{11}$$

where η_{t+i} is the learning rate (a.k.a. step size) and $i = 0, 1, ..., R-1$. After R iterations, FedLEC terminates and returns the latest aggregated global model parameters θ_R as the FedLEC solution, then we have:

Theorem 1. *Let Assumptions 1 to 4 hold and L, μ, σ_m, G be defined therein. Choose $\kappa = \frac{L}{\mu}$, $\gamma = \max\{8\kappa, E\}$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+r)}$. Then FedLEC with full device participation satisfies:*

$$\mathbb{E}[\mathcal{L}(\theta_R)] - \mathcal{L}^* \leq \frac{\kappa}{\gamma + R - 1}\left(\frac{2B}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}||\theta_0 - \theta^*||^2\right) \tag{12}$$

where

$$B = \sum_{m=1}^{M} \mathcal{P}_m^2 \sigma_m^2 + 6L\Gamma + 8(E-1)^2 G^2 \tag{13}$$

If a random set $\mathcal{S}$ of clients is selected randomly without replacement to participate in aggregating, $|\mathcal{S}| = K$, the aggregation step of FedLEC performs $\theta_r \leftarrow \frac{N}{K}\sum_{m \in \mathcal{S}} \mathcal{P}_m \theta^m_r$. By converting the FL objective in Equation (4) as: $\mathcal{L}(\theta) = \frac{1}{M}\sum_{m=1}^{M} \tilde{\mathcal{L}}_m(\theta)$, where $\tilde{\mathcal{L}}_m(\theta) = \mathcal{P}_m M \mathcal{L}_m(\theta)$, we have:

Theorem 2. *Let Assumptions 1 to 5 hold and define $\tilde{C} = \frac{M-K}{M-1}\frac{4}{K}E^2\tilde{G}^2$, then*

$$\mathbb{E}[\mathcal{L}(\theta_R)] - \mathcal{L}^* \leq \frac{\kappa}{\gamma + R - 1}\left(\frac{2(\tilde{B} + \tilde{C})}{\tilde{\mu}} + \frac{\tilde{\mu}\gamma}{2}\mathbb{E}||\theta_0 - \theta^*||^2\right) \tag{14}$$

where $\tilde{B}$ is similar to Equation (13), but L, μ, σ, G are replaced by $\tilde{L} \triangleq \nu L, \tilde{\mu} \triangleq \varsigma\mu, \tilde{\sigma} \triangleq \sqrt{\nu}\sigma$, and $\tilde{G} \triangleq \sqrt{\nu}G$, respectively. Here, $\nu = M \cdot \max_m \mathcal{P}_m$ and $\varsigma = M \cdot \min_m \mathcal{P}_m$

Since $||\theta_0 - \theta^*||^2 \leq \frac{4}{\mu^2}G^2$ for μ-strongly convex $\mathcal{L}$, the dominating term in Equation (14) is:

$$\mathcal{O}\left(\frac{\sum_{m=1}^{N} \mathcal{P}_m^2\sigma_m^2 + L\Gamma + (1 + \frac{1}{K})E^2G^2 + \gamma G^2}{\mu R}\right) \tag{15}$$

Given the pre-defined settings in Assumptions 1 to 4, we conclude that FedLEC will converge with $\mathcal{O}(\frac{1}{R})$ rate in non-IID FL settings like label skewness according to Equation (15).

4.4 Privacy Protection

FedLEC inherits the privacy-preserving merit of the FL framework, which allows each client to preserve private data locally and significantly reduce the risk of privacy leakage. To further handle the potential privacy violation when uploading model parameters to the server, we propose integrating the local differential privacy strategy [El Ouadrhiri and Abdelhadi, 2022] into FedLEC. Specifically, we incorporate a zero-mean Laplacian noise to the SNN model parameters before it is uploaded to the server:

$$\theta^m = \theta^m + \text{Laplacian}(0, \delta) \tag{16}$$

where δ is the noise intensity. Hence, one cannot easily obtain the updated data by monitoring the parameter change, and the privacy protection ability is better as δ increases.

Dataset	Partition	FedAvg	FedProx	Scaffold	FedNova	FedLEC
cifar10	IID	83.91	82.53	**85.18**	79.89	–
	$p \sim Dir(0.05)$	35.80	32.16	29.42	27.37	(↑ 11.07) **46.87**
	$p \sim Dir(0.1)$	53.45	51.76	54.08	47.69	(↑ 13.56) **67.64**
	$\#cnum = 2$	29.60	24.64	34.65	25.61	(↑ 6.51) **41.16**
	$\#cnum = 4$	54.24	56.94	64.28	53.82	(↑ 4.96) **69.24**
cifar100	IID	52.78	46.17	**54.56**	33.37	–
	$p \sim Dir(0.05)$	24.11	25.01	26.11	13.90	(↑ 14.46) **40.57**
	$p \sim Dir(0.1)$	29.29	29.45	33.06	15.51	(↑ 9.67) **42.73**
	$\#cnum = 20$	20.80	22.28	23.01	13.98	(↑ 14.24) **37.25**
	$\#cnum = 40$	32.85	34.41	37.23	20.85	(↑ 9.84) **47.07**
svhn	IID	93.94	93.28	**94.26**	91.03	–
	$p \sim Dir(0.05)$	44.53	46.80	40.24	33.87	(↑ 12.06) **58.86**
	$p \sim Dir(0.1)$	66.04	68.26	66.48	67.82	(↑ 14.70) **82.96**
	$\#cnum = 2$	39.42	31.01	32.00	37.33	(↑ 9.87) **49.29**
	$\#cnum = 4$	65.61	69.87	72.66	69.02	(↑ 9.44) **82.10**

Table 1: Test accuracy (%) of different federated SNN learning approaches w.r.t label skews. The IID results are set as a reference, indicating how extreme label skews affect final accuracy. Typically, The skewness of $\#cnum = 20/40$ for 100-labels is equivalent to $\#cnum = 2/4$ for 10-labels. **Bold** value is the best result across all FL algorithms, while underline value is the second-best. We run *three* trials and report the mean top-1 accuracy.

5 Experiments

5.1 Experimental Settings

In this study, we use **four** FL algorithms designed for tackling conventional data heterogeneity, i.e., FedAvg, FedProx [Li *et al.*, 2020], FedNova [Wang *et al.*, 2020], and Scaffold [Karimireddy *et al.*, 2020]), and **four** extra FL algorithms proposed to tackle the label skewness, i.e., FedLC [Zhang *et al.*, 2022], FedRS [Li and Zhan, 2021], FLea [Xia *et al.*, 2024], and FedConcat [Diao *et al.*, 2024], as compared baselines to demonstrate the effectiveness of FedLEC. All FL algorithms adopt *S-VGG9* [Venkatesha *et al.*, 2021] as the default backbone SNN model.

We uniformly execute 50 communication rounds, selecting 20% parts from 10 clients to train their models for 10 local epochs per round with a batch size of 64 as default. We use the Adam optimizer for all trials with a learning rate 0.001. All experimental trials are implemented on NVIDIA GeForce RTX 4090 GPUs.

5.2 Performance Evaluation

Overall Accuracy. Table 1 presents the accuracy performance of FedLEC compared to conventional FL algorithms tailored for data heterogeneity. The results demonstrate the clear superiority of FedLEC in handling extreme label skew conditions. Specifically, in *cifar10* dataset, FedLEC reaches an average improvement of about 9.03%, while in *svhn*, the average accuracy increment rate is 11.52%. When the task extends to more complex classification tasks such as *cifar100*, FedLEC similarly exhibits notable performance improvements, highlighting its potential for adaptation to complex and large-scale tasks. Additionally, FedLEC demonstrates potential advantages in handling distribution-based label skews, as these skews do not result in the absence of a significant portion of labels. Consequently, they do not induce severe information loss, unlike quantity-based skews.

Table 2 further provides a performance comparison between FedLEC and several recently proposed FL algorithms

Dataset	Partition	FedLC	FedRS	FLea	FedConcat	FedLEC
cifar10	$\#cnum = 2$	23.96	24.28	<u>32.66</u>	32.64	(↑ 8.50) **41.16**
	$p \sim Dir(0.05)$	32.94	<u>33.49</u>	28.16	30.43	(↑ 13.38) **46.87**
cifar100	$\#cnum = 20$	19.88	8.01	<u>28.57</u>	25.78	(↑ 8.68) **37.25**
	$p \sim Dir(0.05)$	21.96	12.82	<u>26.62</u>	23.15	(↑ 13.95) **40.57**
svhn	$\#cnum = 2$	31.69	31.71	<u>33.46</u>	32.84	(↑ 15.83) **49.29**
	$p \sim Dir(0.05)$	29.97	39.10	42.67	<u>45.63</u>	(↑ 13.23) **58.86**

Table 2: Test accuracy (%) of FedLEC compared with the other baselines specially designed for addressing extreme label skews, denoted in the same way as Table 1.

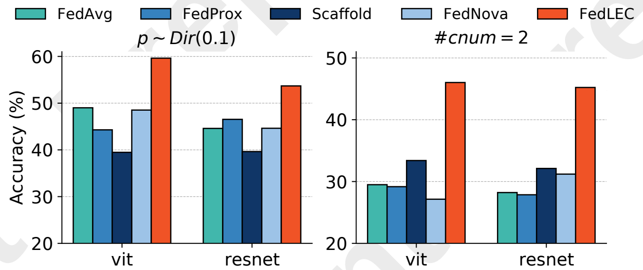

Figure 4: Accuracy performance of different models when varying various FL algorithms under different label skews.

specifically designed to mitigate label skewness. Observations reveal that these algorithms are ineffective in federated SNN learning settings, exhibiting an average accuracy that is 12.26% lower than that of FedLEC.

The superiority of FedLEC may stem from the **fundamental differences** in training paradigms between ANNs and SNNs. BPTT with surrogate functions induces the intrinsic gradient bias accumulated by layer-wise back-propagation [Deng *et al.*, 2023]. Extreme label skewness aggravates this bias, leading to drifting errors in the features extracted by the SNN model. In particular, algorithms like FLea and FedConcat heavily rely on features extracted by local models to enhance resistance to label skewness. However, this dependency may backfire, further amplifying the occurrence of errors. The distillation loss $\mathcal{L}_d$ can effectively mitigate excessive gradient drift. Since the predictions of the previous global model retain partial information about the correct optimization direction, they can serve as a reference during the current phase of local model training, helping to prevent over-fitting. Hence, it is reasonable that the performance of FedLEC surpasses the other baselines by an average of approximately 11.33%.

Results on Other SNN Architectures. To assess the generalization capability of FedLEC, we incorporate more advanced SNN models, such as *MS-ResNet* [Hu *et al.*, 2024] and *Meta-Spikeformer* [Yao *et al.*, 2024], for training in an FL setting under various label skewness scenarios. For brevity, we simplify *MS-ResNet* as *resnet* and *Meta-Spikeformer* as *vit*. As depicted in Figure 4, FedLEC achieves accuracy improvements of 10.13% and 11.62% for *resnet* and *vit*, respectively, demonstrating that the proposed local objective remains effective even for these complex SNN architectures.

Results on Event-Based Datasets. We also apply FedLEC to event-based image classification tasks, a representative application scenario for SNNs, to evaluate its generalized effectiveness across a broader range of scenarios. Figure 5 shows that while FedLEC continues to exhibit superior performance, the performance gaps between FedLEC and other algorithms are notably smaller than those observed in static image classification tasks, with an average improvement of only approximately 4%. In particular, the classification task on the event-based dataset is more complicated since the features are sparse and chaotic. Hence, achieving a well-performing global model capable of capturing accurate patterns becomes challenging with highly skewed datasets, which limits the performance improvement of FedLEC.

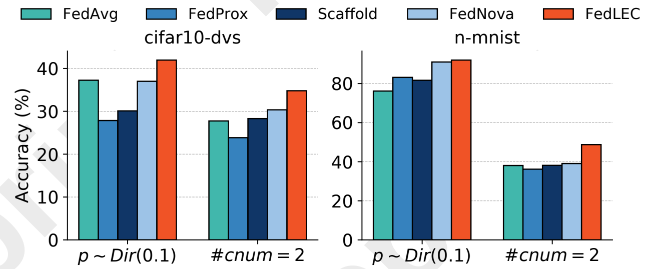

Figure 5: Accuracy performance of various federated SNN learning algorithms across different data scenarios under different skewness.

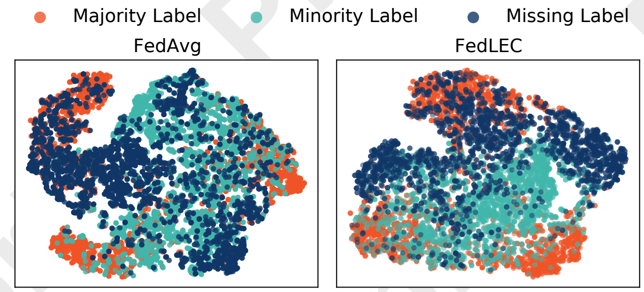

Figure 6: T-SNE visualizations on majority, minority, and missing labels. **Left:** For FedAvg, the samples from the minority and missing labels are mixed and indistinguishable. **Right:** For our method, the data from minority and missing labels can be distinguished well.

5.3 Efficiency Evaluation

T-SNE Visualization. We provide a visual analysis that examines the differences in the sample features extracted by the same models following local updates between FedLEC and FedAvg, aiming to demonstrate FedLEC's effectiveness. We first randomly select a client with a local data shard where $|D_{\mathcal{M}}| = 0$, $|D_{\mathcal{K}}| = 249$, $|D_{\mathcal{J}}| = 8437$. After the well-trained global models are locally trained in this client using FedAvg and FedLEC with one epoch, we feed them with the same test data and leverage the corresponding output features for visualization [Van der Maaten and Hinton, 2008]. As depicted in Figure 6, the test samples, especially those from minority (i.e., green samples) and missing (i.e., dark blue samples) labels, are mixed and difficult to distinguish. However, FedLEC can alleviate this issue and learn more discriminative features, reducing classification ambiguity when making inferences on samples from missing and minority labels.

Ablation Study. The results in Table 3 quantify the contribution of each local loss term in FedLEC. From these results, we conclude that both terms significantly contribute to

$\mathcal{L}_c$	$\mathcal{L}_d$	*cifar10*		*svhn*	
		$p \sim Dir(0.15)$	$\#cnum = 3$	$p \sim Dir(0.15)$	$\#cnum = 3$
✗	✗	63.35	32.24	70.29	46.36
✓	✗	(↑ 0.65) 64.00	(↑ 3.06) 35.30	(↑ 2.89) 73.18	(↑ 0.42) 46.78
✗	✓	(↑ 5.45) 68.80	(↑ 24.6) 56.79	(↑ 13.7) 84.02	(↑ 26.7) 73.08
✓	✓	(↑ 6.04) **69.39**	(↑ 25.1) **57.30**	(↑ 14.0) **84.24**	(↑ 30.1) **76.49**

Table 3: Impact of each loss component on Accuracy (%).

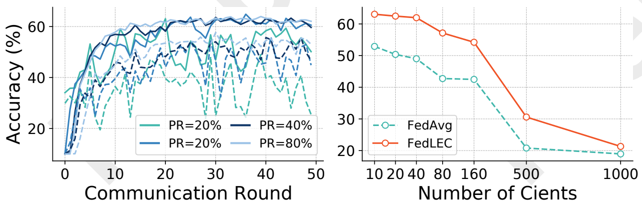

Figure 7: Accuracy patterns of SNN-based FedAvg (dashed line) and FedLEC (solid line) when the participation rate (PR) and the total number of local clients vary under the label skew $p \sim Dir(0.1)$.

the final accuracy, with $\mathcal{L}_d$ being the dominant factor, leading to an average improvement of 17.61%. In addition, the combination of both loss terms further enhances the federated SNN learning performance under extreme label skew, yielding an additional 1.20% improvement. This improvement is attributed to $\mathcal{L}_d$ mitigating gradient biases and conveying the global data distribution, while $\mathcal{L}_c$ helps alleviate the distribution imbalance.

5.4 Hyper-Parameter Analysis

Client Participation Rate. In reality, Not all the clients will participate in the entire training process. As the left part of Figure 7 illustrates, we simulate this scenario by setting the sample fraction from 20% to 80%. The FedAvg training curves are less stable than those of FedLEC due to the highly skewed data. The local gradient distributions vary among communication rounds, distorting the correct global model updating direction per round.

Total Client Number. We study the effect of the number of clients on FedLEC as shown in the right part of Figure 7. We observe that the overall accuracy trends of both FL algorithms tend to decline as the number of clients increases. FedLEC consistently outperforms FedAvg across all large-scale settings, achieving an average improvement of 12.27%. Although increasing the number of clients reduces the volume of local data per client, which amplifies the risk of biased training, FedLEC effectively mitigates this issue.

Time Steps. The number of time steps dictates the complexity of federated SNN learning. We systematically vary this parameter to assess its impact on FedLEC's performance. Observations in the left part of Figure 8 illustrate a continual improvement in performance for FedLEC as time steps increase from 42.74% to 47.55%, and the improvement is marginal when the time step increases to 7, as the membrane potential of LIF neurons requires several time steps to stabilize [Anumasa *et al.*, 2024] after initialization, allowing the SNN model to generate stable spike trains effectively.

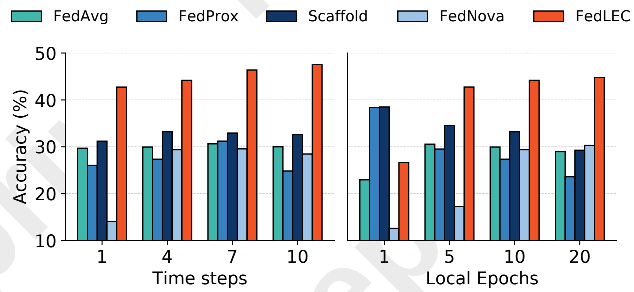

Figure 8: Accuracy patterns of different federated SNN learning algorithms when the time steps and local training epochs vary under the label skew $\#cnum = 2$.

Intensity δ	0.7	0.5	0.3	0.1	0
Accuracy (%)	19.61	19.49	31.06	61.41	63.70

Table 4: Results of applying local differential privacy technique into FedLEC with various noise intensity δ.

Local Epochs. We vary the local epochs from {1, 5, 10, 20} and report the final accuracy under specific label skew in the right part of Figure 8. On the one hand, all FL algorithms are sensitive to the number of local epochs. When the number of local epochs increases from 1 to 5, FedLEC achieves a more substantial accuracy improvement of 16.10% compared to other algorithms, as incorporating regularization requires additional epochs for convergence. On the other hand, the optimal local epoch number varies across FL algorithms under a specific label skew. When local epochs reach 10, the accuracy improvement rate turns marginal for FedLEC and FedNova, whereas algorithms like FedProx and Scaffold might prefer fewer local epochs.

Privacy Protection Intensity. We evaluate the performance of our privacy protection enhanced FedLEC with the local differential privacy strategy. In Particular, we set the noise intensity $\delta = [0, 0.1, 0.3, 0.5, 0.7]$ and the corresponding experimental results are shown in Table 4. We can see that the performance declines as the noise intensity δ grows, while the performance drop is slight if δ is not too large. When δ is larger than 0.5, the final accuracy no longer continues to decrease. Hence, a moderate strength of δ, such as 0.1, is desirable to balance FedLEC accuracy and privacy protection.

6 Conclusion

This study addresses the critical challenge of accuracy deterioration in federated SNN learning caused by highly label-skewed data. The proposed FedLEC addresses the issue of imbalanced local data distribution by employing a calibrated objective function and mitigates the learning bias exacerbated by missing labels through knowledge distillation for inter-client label alignment. Extensive experiments on diverse datasets demonstrate that FedLEC significantly outperforms various state-of-the-art FL approaches in most label-skewed scenarios. These results highlight the potential of FedLEC for real-world applications, particularly in energy-constrained and data-heterogeneous edge environments.

Acknowledgments

The work of this paper is supported by the National Key Research and Development Program of China under Grant 2022YFB4500100, the National Natural Science Foundation of China under Grant 62125206, the Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020014, the National Key Research and Development Program of China No. 2024YDLN0005, and the Regional Innovation and Development Joint Fund of the National Natural Science Foundation of China No. U22A6001.

References

[Anumasa *et al.*, 2024] Srinivas Anumasa, Bhaskar Mukhoty, Velibor Bojkovic, Giulia De Masi, Huan Xiong, and Bin Gu. Enhancing training of spiking neural network with stochastic latency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10900–10908, 2024.

[Aouedi and Kandaraj, 2024] Ons Aouedi and Piamrat Kandaraj. Surfs: Sustainable intrusion detection with hierarchical federated spiking neural networks. *ICC 2024*, 2024.

[Aouedi *et al.*, 2023] Ons Aouedi, Kandaraj Piamrat, and Mario Sûdholt. Hfedsnn: Efficient hierarchical federated learning using spiking neural networks. In *Proceedings of the Int'l ACM Symposium on Mobility Management and Wireless Access*, pages 53–60, 2023.

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

[Di *et al.*, 2024] Yu Di, Du Xin, Linshan Jiang, Wentao Tong, and Shuiguang Deng. Ec-snn: Splitting deep spiking neural networks for edge devices. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence (2024).*, volume 271507864, 2024.

[Diao *et al.*, 2024] Yiqun Diao, Qinbin Li, and Bingsheng He. Exploiting label skews in federated learning with model concatenation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11784–11792, 2024.

[El Ouadrhiri and Abdelhadi, 2022] Ahmed El Ouadrhiri and Ahmed Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE access*, 10:22359–22380, 2022.

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

[Huang, 2005] Jonathan Huang. Maximum likelihood estimation of dirichlet distribution parameters. *CMU Technique report*, 76, 2005.

[Karimireddy *et al.*, 2020] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[Li and Zhan, 2021] Xin-Chun Li and De-Chuan Zhan. Fedrs: Federated learning with restricted softmax for label distribution non-iid data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 995–1005, 2021.

[Li *et al.*, 2019] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[Li *et al.*, 2022] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE, 2022.

[Lim *et al.*, 2020] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE communications surveys & tutorials*, 22(3):2031–2063, 2020.

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

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Menon *et al.*, 2021] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

[Su *et al.*, 2023] Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6555–6565, 2023.

[Tumpa *et al.*, 2023] Sadia Anjum Tumpa, Sonali Singh, Md Fahim Faysal Khan, Mahmut Tylan Kandemir, Vijaykrishnan Narayanan, and Chita R Das. Federated learning with spiking neural networks in heterogeneous systems. In *2023 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 1–6. IEEE, 2023.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Venkatesha *et al.*, 2021] Yeshwanth Venkatesha, Youngeun Kim, Leandros Tassiulas, and Priyadarshini Panda. Federated learning with spiking neural networks. *IEEE Transactions on Signal Processing*, 69:6183–6194, 2021.

[Wang *et al.*, 2020] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

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

[Wu *et al.*, 2019] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1311–1318, 2019.

[Xia *et al.*, 2024] Tong Xia, Abhirup Ghosh, Xinchi Qiu, and Cecilia Mascolo. Flea: Addressing data scarcity and label skew in federated learning via privacy-preserving feature augmentation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3484–3494, 2024.

[Xie *et al.*, 2022] Kan Xie, Zhe Zhang, Bo Li, Jiawen Kang, Dusit Niyato, Shengli Xie, and Yi Wu. Efficient federated learning with spike neural networks for traffic sign recognition. *IEEE Transactions on Vehicular Technology*, 71(9):9980–9992, 2022.

[Yao *et al.*, 2024] 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. In *The Twelfth International Conference on Learning Representations*, 2024.

[Yurochkin *et al.*, 2019] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR, 2019.

[Zhang *et al.*, 2022] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pages 26311–26329. PMLR, 2022.

[Zhang *et al.*, 2024] Mengjun Zhang, Bin Li, Hongfu Liu, and Chenglin Zhao. Federated learning for radar gesture recognition based on spike timing dependent plasticity. *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[Zuo *et al.*, 2024] Lin Zuo, Yongqi Ding, Mengmeng Jing, Kunshan Yang, and Yunqian Yu. Self-distillation learning based on temporal-spatial consistency for spiking neural networks. *arXiv preprint arXiv:2406.07862*, 2024.