Global Balanced Experts for Federated Long-Tailed Learning

Yaopei Zeng[1,*], Lei Liu[1,*], Li Liu[2,†], Li Shen[3], Shaoguo Liu[4], Baoyuan Wu[1,†]
[1]School of Data Science, Shenzhen Research Institute of Big Data,
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China
[2] Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, China
[3]JD Explore Academy, [4]Alibaba Group

Abstract

*Federated learning (FL) is a prevalent distributed machine learning approach that enables collaborative training of a global model across multiple devices without sharing local data. However, the presence of long-tailed data can negatively deteriorate the model's performance in real-world FL applications. Moreover, existing re-balance strategies are less effective for the federated long-tailed issue when directly utilizing local label distribution as the class prior at the clients' side. To this end, we propose a novel **G**lobal **B**alanced **M**ulti-**E**xpert (GBME) framework to optimize a balanced global objective, which does not require additional information beyond the standard FL pipeline. In particular, a proxy is derived from the accumulated gradients uploaded by the clients after local training, and is shared by all clients as the class prior for re-balance training. Such a proxy can also guide the client grouping to train a multi-expert model, where the knowledge from different clients can be aggregated via the ensemble of different experts corresponding to different client groups. To further strengthen the privacy-preserving ability, we present a GBME-p algorithm with a theoretical guarantee to prevent privacy leakage from the proxy. Extensive experiments on long-tailed decentralized datasets demonstrate the effectiveness of GBME and GBME-p, both of which show superior performance to state-of-the-art methods. The code is available at here.*

1. Introduction

Federated Learning (FL) is a collaborative training method to develop a global model by utilizing decentralized data from multiple clients [21]. It enables knowledge aggregation over disparate data sources while mitigating privacy risks for individual clients. However, the model's performance of a FL system can be severely deteriorated in the presence of long-tailed data distribution, which is an ubiquitous problem in various realistic scenarios [13, 36], such as medical applications [17], personal information protection [44] and autonomous vehicles [26].

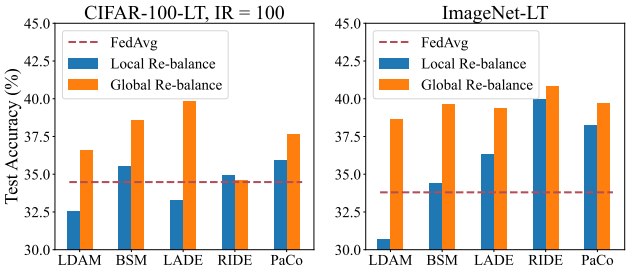

Figure 1: Global re-balance *VS.* Local re-balance[1]. X-axis denotes several classical re-balance strategies. Global re-balance significantly outperforms local re-balance in the FL setting with long-tailed data, motivating us to optimize a global balanced objective.

Importantly, it is extremely difficult to learn a balanced global model in the FL setting with long-tailed data [34], especially for the minority classes [30]. In the concrete, due to data heterogeneity, there may exist a large divergence among the imbalanced distributions of different clients, *e.g.*, different local datasets have different imbalance ratios or minority classes (visualized in Section 9 in **Supplementary Material**). Additionally, during the standard FL training, partial client selection may randomly drop some minority samples at each communication round, further decreasing the model performance on minority classes. Therefore, the long-tailed issue is more challenging in the FL scenarios.

Several techniques have been proposed to tackle the federated long-tailed problem, such as loss re-weighting [30, 34], client clustering [6], and client selection [39].

* indicates equal contribution.
† denotes corresponding author (avrillliu@hkust-gz.edu.cn, wubaoyuan@cuhk.edu.cn).

[1]Global re-balance means that the re-balance strategies adopt global label distribution as the class prior, while local re-balance utilizes local label distribution as the class prior.

However, it is generally assumed that some sensitive information is accessible to the server, *e.g.*, a balanced mini-dataset [6, 34] or learnable hyper-parameters of clients [30], which may not be available in realistic applications. Besides, most of them focus on datasets with few classes (*e.g.*, ten or twenty), and their effectiveness diminishes on large-scale imbalanced datasets with a higher number of classes [20, 43]. Furthermore, we report the performance of several classical re-balance methods [2, 25, 12, 35, 4] under the federated long-tailed setting in Figure 1. It is observed that the performance improvement is limited compared with FedAvg [21] when taking local label distribution as the class prior of re-balancing strategies (*i.e.*, local re-balance).

Aiming to tackle the above problem, this work further explores the effectiveness of existing class-prior based re-balance algorithms for federated long-tailed learning. Experimentally, as indicated by Figure 1, deploying these algorithms with global re-balance yields higher accuracy than that with local re-balance[1]. The main reason arises from the optimization objective gap between these two re-balance strategies, where the global one provides a consistent objective with the centralized balancing training (introduced in Section 3.1). However, global re-balance requires clients to upload local label distributions to obtain global label distribution, thus increasing the risk of privacy leakage [34].

To get rid of those constraints, we propose a **G**lobal **B**alanced **M**ulti-**E**xpert (GBME) framework to deal with the federated long-tailed issue without requiring additional information beyond the standard FL. Specifically, we derive a local proxy from the accumulated gradients of the clients after local training rather than from the local label distribution, and then a global proxy is formulated as the class prior for re-balance algorithms by integrating local proxies of each client. Based on the cosine similarity between the local and global proxy, clients can be divided into different groups corresponding to different experts in a multi-expert model. During the local training of a client, the corresponding expert is trainable to learn balanced knowledge using the global proxy as the prior, while other experts are frozen to maintain the knowledge learned from other groups. Using a multiple selection strategy, a client can implicitly interact with other groups in an ensemble manner to aggregate balanced knowledge learned from different groups. To further improve the privacy-preserving ability, we present a GBME-p algorithm based on the differential privacy (DP) [7] to prevent the privacy leakage of local label distributions. Concretely, the random Gaussian noises are added to the weights of the final fully connected (FC) layer for local proxy computation at the clients' side before uploading. The overall GBME framework is illustrated in Figure 2. In summary, the key contributions of this work are as follows.

(i) We experimentally and theoretically explore the effectiveness of existing class-prior based re-balance algorithms in federated long-tailed learning. It is demonstrated that there is a mismatch between the optimization objectives of local and global re-balance strategies, where global re-balance performs better than the local one on the imbalanced decentralized data.

(ii) We propose a GBME framework to achieve global balanced training, where a proxy is designed as the class prior for re-balancing algorithms without requiring additional private information. The clients are divided into multiple groups to collaboratively train a multi-expert model, where the knowledge from different groups can be aggregated in an ensemble manner.

(iii) To enhance the privacy-preserving ability, we present a GBME-p algorithm with a theoretical guarantee to prevent the privacy leakage of local label distributions, where the Gaussian noises are added to the weights of the last FC layer at the clients' side before uploading.

(iv) The experiments on multiple benchmark datasets demonstrate that GBME without requiring additional private information can significantly outperform previous state-of-the-art (SOTA) methods. Besides, GBME-p can still achieve superior performance under the protection of the differential privacy.

2. Related Work

Federated Learning. FL [21] is a learning framework to train a global model on distributed data of multiple clients with privacy protection. One of the most important challenges is data heterogeneity. Many previous studies focused on this problem [15, 18, 31, 27, 33] with the assumption of a perfectly balanced global dataset (all local data). Recent works [6, 39, 34, 30] proposed to handle class imbalance issue in FL. For example, CReFF [28] deal with federated long-tailed data inspired by [14]. However, these methods usually required additional private information except for model parameters of the clients with the privacy concerns, *e.g.*, CReFF [28] requires feature gradients of clients' data. Moreover, they only focused on datasets with a few classes, and their effectiveness may diminish on the large-scale imbalanced datasets with a larger amount of classes [30, 45].

Long-tailed Learning. Real-world data often exhibits a long-tailed distribution, where the majority classes have massive samples and the minority classes only have a few samples [43]. Many re-balance strategies are proposed to address such imbalance issues. Data re-sampling [3, 10, 14] is a common type, such as over-sampling the minority samples [29, 14] or under-sampling [10] the majority samples. Another scheme to learn a balanced model is loss re-weighting [32, 5]. Generally, these methods tend to

give a large training loss for the minority samples. Recent studies mainly focused on a good representation space to improve the generalization ability. PaCo [4] introduced a contrastive learning method over the long-tailed dataset. Ensemble learning is also effective in long-tailed learning [46, 37, 35, 42, 1]. Although these re-balance strategies worked well on the centralized imbalance datasets, it remains a question whether they are useful in federated long-tailed learning. In this work, we theoretically and experimentally explore this issue and propose a novel algorithm to achieve a global balance training with existing re-balance strategies for federated long-tailed learning.

3. GBME Learning Method

Problem Formulation. We discuss a typical FL setting with total K clients. A distinct data source $\mathcal{D}^k$ containing N^k samples is held by client C_k. The global dataset is defined as $\mathcal{D} = \bigcup_{k\in[K]} \mathcal{D}^k$. Considering a S-class classification task, $(\boldsymbol{x}, y) \in \mathcal{D}$ is a training sample, where $\boldsymbol{x}$ is an image in the input space $\mathcal{X}$ and y is its corresponding label. Assume $\mathcal{D}$ follows a long-tailed distribution, *i.e.*, the sample size is exponentially distributed *w.r.t.* class index. The global imbalance ratio (IR) is defined as the ratio between the largest and smallest class volumes. Typically, FL aims at learning a single shared model and optimizing the global objective as the aggregation of the local objectives:

$$\min_{\theta\in\mathbb{R}^d} \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta), \ \ \textit{where} \ \ F_k(\theta) = \frac{1}{n_k} \sum_{(\boldsymbol{x},y)\in\mathcal{D}^k} f_\theta(\boldsymbol{x}, y), \quad (1)$$

where the model is parameterized by θ and f is the loss function. In the presence of class imbalance, the above formulation suffers from both imbalanced and heterogeneous data, which easily produces a model that performs poorly on the minority classes. We summary the main challenges of federated long-tailed learning as follows.

Challenges. (1) Local datasets may exhibit varying imbalanced distributions due to the distinct data sources, and significantly differ from the global imbalanced distribution, resulting in more severe heterogeneity. We visualize such issue in Section 9 in **Supplementary Material**. (2) Partial client selection for FL communications may drop some minority samples, further aggravating the imbalance issue because of insufficient minority data. As a theoretical motivation, we consider the following configuration to investigate the effectiveness of class prior based re-balance techniques, which inspires us to optimize a global balanced objective instead of a local one for federated long-tailed learning.

3.1. Theoretical Motivation

We consider a binary classification problem where the ground truth is either positive (y^+) or negative (y^-). Under the heterogeneously imbalanced setting, given two clients C_0 and C_1, we assume that client C_0 accesses n_0^+ positives and n_0^- negatives, while client C_1 accesses n_1^+ positives and n_1^- negatives. Without loss of generality, we consider a simple loss re-weighting strategy that takes the inverse of the proportion of each class as the weight for this class, *i.e.*, n_0/n_0^+ for the positive class and n_0/n_0^- for the negative class on client C_0. Similarly, client C_1 uses n_1/n_1^+ for the positive class and n_1/n_1^- for the negative class. The global re-balance strategy takes the inverse of the global label distribution as the class weights, *i.e.*, $(n_0+n_1)/(n_0^+ + n_1^+)$ for positive class and $(n_0 + n_1)/(n_0^- + n_1^-)$ for negative class. We denote the global objectives of the global and local re-balance strategies as $G_g(\theta)$ and $G_l(\theta)$, respectively. Then we can measure the difference between $G_g(\theta)$ and $G_l(\theta)$.

Lemma 1 *Using the global re-balance strategy, the global objective yields the same form as the objective of the re-balance methods on the centralized dataset:*

$$G_g(\theta) = \frac{1}{n^+} \sum_{(\boldsymbol{x},y^+)\in\mathcal{D}} f_\theta(\boldsymbol{x}, y^+) + \frac{1}{n^-} \sum_{(\boldsymbol{x},y^-)\in\mathcal{D}} f_\theta(\boldsymbol{x}, y^-). \quad (2)$$

where $n^+ = n_0^+ + n_1^+$ *and* $n^- = n_0^- + n_1^-$.

Theorem 1 *Under the above setting, let $\mathcal{E}$ be the estimation of global label distribution. There exists a group of re-balance weights e derived from $\mathcal{E}$, whose global objective G_e satisfies $G_g(\theta) \le G_e(\theta) < G_l(\theta)$. Then the optimization objective gap $\Delta = G_l(\theta) - G_g(\theta)$ can be written as:*

$$\begin{aligned}\Delta = &\frac{n_1(n_0^+)^2 + n_0(n_1^+)^2}{n_0^+ n_1^+ (n_0+n_1)(n_0^+ + n_1^+)} \sum_{(\boldsymbol{x},y^+)\in\mathcal{D}} f_\theta(\boldsymbol{x}, y^+) \\ &+ \frac{n_1(n_0^-)^2 + n_0(n_1^-)^2}{n_0^- n_1^- (n_0+n_1)(n_0^- + n_1^-)} \sum_{(\boldsymbol{x},y^-)\in\mathcal{D}} f_\theta(\boldsymbol{x}, y^-).\end{aligned} \quad (3)$$

Interpretation. The above theorem illustrates the following points: (1) The local re-balance strategy optimizes a larger objective than the global one, which yields the same objective function as the re-balance strategies on the centralized dataset. (2) There exists a re-balance strategy derived from the global perspective, which exhibits a smaller objective than the local re-balance strategies.

3.2. Global Proxy Information

Our theoretical results show that the global re-balance strategy is more suitable in the federated long-tailed setting. However, global label distribution is unavailable due to the privacy concerns of FL. To overcome this drawback, we propose a proxy called Global Proxy Information (GPI) derived from the accumulated gradients of the clients, which is inspired by an empirical observation [14]: in the FC layer,

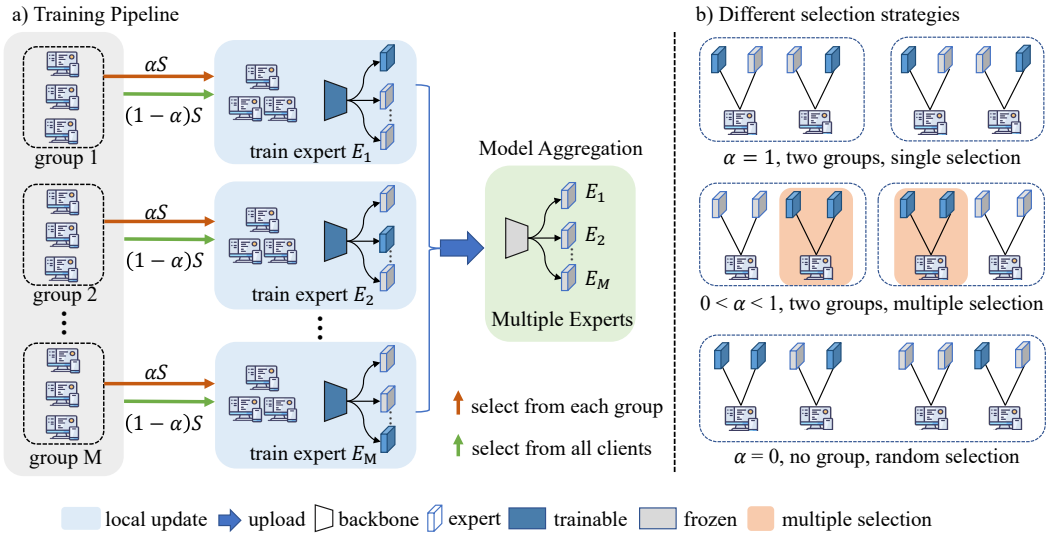

Figure 2: Multi-Expert Learning and Different Selection Strategies. **(a) Training Pipeline.** We first divide all clients into different groups according to the similarity between their LPI and GPI. Then we select diverse clients for training different experts in an ensemble manner. **(b) Client selection with different α values.** If $\alpha = 1$, most clients are only selected one time to update the corresponding expert of their groups. If $0 < \alpha \leq 1$, a client may be selected multiple times to update multiple experts corresponding to their groups and other groups, where different groups can interact with each other via such selection. $\alpha = 0$ indicates that a client is randomly selected to update one of the experts.

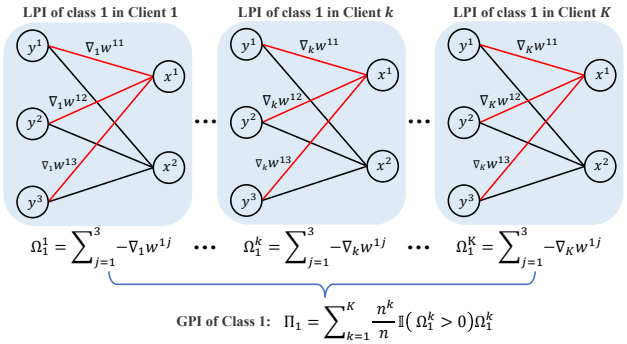

Figure 3: Calculations of LPI and GPI.

the weight norms of majority classes are generally larger than that of minorities.

Definition 1 *(Local Proxy Information). For local training on $\mathcal{D}^k$, $\nabla_k w^{ij}$ is the gradient of the weight w^{ij} connecting j-th input with i-th output in the FC layer, where the input dimension of the FC layer is H. Local proxy information (LPI) of class i on $\mathcal{D}^k$ is defined as the gradient magnitude associated with i-th output of the FC layer:*

$$\Omega_i^k = \sum_{j=1}^{H} -\nabla_k w^{ij}, \text{ where } \nabla_k w^{ij} = \frac{\partial F_k(\theta)}{\partial w^{ij}}, \tag{4}$$

where Ω_i^k is i-th element in the local proxy information vector Ω^k of the client C_k.

Definition 2 *(Global Proxy Information). The global proxy information (GPI) of class i is defined as the weighted summation of the local proxy information:*

$$\Pi_i := \sum_{k=1}^{K} \frac{n^k}{n} \mathbb{I}(\Omega_i^k > 0) \Omega_i^k, \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function and the value is 1 if $\cdot$ is true, 0 otherwise. Π_i is i-th element in the GPI vector Π.

$\mathbb{I}(\cdot)$ is necessary since negative LPI value may result in unstable behavior of GPI. In fact, as shown in Figure 3, the LPI estimation only involves the accumulated gradients of the FC layer, thus the central server receives no additional private information compared with the vanilla FedAvg [21]. GPI can be obtained by aggregating the accumulated gradients of different clients. Then the server broadcasts GPI to each client, where GPI is used as the class prior for existing re-balance strategies during local training. Besides, the clients are required to upload LPI only in the beginning commutation round, then global re-balance training can be conducted in the subsequent FL communication rounds.

Experimentally, as shown in Section 4, we find that previous re-balance strategies with GPI can effectively improve the accuracy of medium-/few-shot classes while reducing the performance drop of many-shot classes.

3.3. Multi-Expert Learning

Guided by our proxy analysis, we utilize a multi-expert architecture to handle the federated long-tailed issue. The global and local models have the same structure, *i.e.*, a shared backbone followed by multiple experts and each expert has individual learnable blocks. The algorithm is illustrated in Figure 2 and Algorithm 1 with the following steps.

Client Grouping. Based on the Section 3.2, we define the cosine similarity between LPI and GPI of client C_k as:

$$\text{Cosine}(\Omega^k, \Pi) = \frac{\Omega^k \cdot \Pi}{\|\Omega^k\| \|\Pi\|}, \tag{6}$$

where $\cdot$ denotes the dot product operation. Then we can obtain the ranked clients by sorting the similarity scores and the clients with closed similarity are divided into the same group. Let hyper-parameter M be the group number, M experts are allocated for each client corresponding these groups respectively, *i.e.*, group P_i corresponds to expert E_i.

Client Selection. In each communication round, two parts of the clients are selected to update the expert E_i: (1) αR clients are randomly selected from the group P_i; (2) $(1-\alpha)R$ clients are randomly selected from other clients. $0 \leq \alpha \leq 1$ is a hyper-parameter to control the client selection, which influences the interactions among different groups. As shown in Figure 2(b), if $\alpha = 1$, there is no interaction among different groups, where each group only updates the corresponding expert. If $\alpha = 0$, by M times bootstrap sampling, i-th selected client is to update E_i individually. To enhance the interaction among different groups, we utilize $0 \leq \alpha \leq 1$ to achieve a multiple selection strategy as shown in Figure 2(b).

Expert Ensemble. The server broadcasts the global model to the selected clients for local training. For each client, the classification loss is calculated based on the average logits of all experts to integrate the knowledge of different groups. Here, we adopt the balanced softmax loss (BSM) [25] due to the effectiveness of logit adjustment for class imbalance issue [24]. Denote the class prior information as π, *e.g.*, local label distribution, global label distribution, and GPI. BSM loss is written as:

$$L_{\text{BSM}} = \frac{1}{n_k} \sum_{(\boldsymbol{x},y)\in\mathcal{D}^k} -y \log s \left(\frac{1}{M} \sum_{i=1}^{M} \mathbf{v}_i\left(\boldsymbol{x}, \theta\right) + \log \pi \right), \tag{7}$$

Algorithm 1: GBME(-p) Framework

```
Initialization: t = 0, and θ_i^(0) = θ^(0), ∀i;
while D^k ∈ {D^1, D^2, ··· , D^K} do
    one optimization round on D^k: θ̂_k^(0);
    GBME-p: ∇_k ŵ^ij = ∇_k w^ij + N_k^(t);
    compute LPI Ω^k for client k by Definition 1;
    upload Ω^k and θ̂_k^(0);
end
compute GPI Π_c for each class c by Definition 2;
client grouping by Eq. (6): {P_1, ··· , P_M};
while t < T do
    broadcast the global model: θ_i^(t) = θ^(t);
    client selection for updating each expert;
    while D^i ∈ {D^1, D^2, ··· , D^R} do
        θ_i^(t+1) = arg min L_BSM(θ_i^(t), D^i)
    end
    global aggregation: θ^(t+1) = Σ_{i=1}^K (n_i/n) θ_i^(t+1);
    t ← t + 1
end
```

where $s(\cdot)$ is the softmax function and $\mathbf{v}_i(\cdot)$ is the output logits of expert E_i. During the local training of a client, only the backbone and selected experts are updated, while other experts are kept frozen. After local updates, the selected clients uploads the updated model parameters and the server aggregates them into the global model with FedAvg [21].

3.4. GBME-p with Privacy Guarantee

In this section, we propose a GBME-p algorithm to address the federated long-tailed problem with privacy protection based on the concept of differential privacy (DP) [7], which provides a theoretical criterion for privacy preservation of distributed learning systems.

Definition 3 *(Differential Privacy [7]). A randomized algorithm $\mathcal{M}: \mathcal{P} \to \mathcal{R}$ with domain $\mathcal{P}$ and range $\mathcal{R}$ is (ϵ, δ)-DP (ϵ is privacy budget and δ is failure probability), if for all measurable sets $\mathcal{R}' \subseteq \mathcal{R}$ for any two adjacent databases $\mathcal{D}_i, \mathcal{D}_i' \in \mathcal{P}$:*

$$\Pr[\mathcal{M}(\mathcal{D}_i) \in \mathcal{R}'] \leq e^{\epsilon} \Pr[\mathcal{M}(\mathcal{D}_i') \in \mathcal{R}'] + \delta. \tag{8}$$

Definition 4 *(Gaussian Mechanism). Given any function $\mathcal{M}: \mathcal{P} \to \mathcal{R}$, the Gaussian mechanism is defined as: $\mathcal{M}_g(\mathcal{D}, \sigma) = \mathcal{M}(\mathcal{D}) + \mathcal{N}$, where $\mathcal{N}$ is a random variable drawn from the Gaussian(σ^2) distribution, where σ is the standard derivation.*

Theorem 2 *For $c^2 > 2\ln(1.25/\delta)$, the Gaussian Mechanism with parameter $\sigma \geq c\Delta/\epsilon$ is (ϵ, δ)-DP with arbitrary $\epsilon \in (0, 1)$. c is influenced by δ to adjust σ.*

GBME-p. Inspired by the above DP mechanism, we propose an algorithm named GBME-p that incorporates the privacy protection ability into the GBME. Concretely, we adopt the Gaussian mechanism for the LPI estimation, *i.e.*, the Gaussian noises parameterized by privacy-related parameters (*e.g.*, ϵ, Δ) are added to the weights of the final FC layer. Algorithm 1 outlines the GBME(-p) algorithm for training a balanced model with (ϵ, δ)-DP requirement.

In fact, (ϵ, δ)-DP ensures that LPI is protected against potential differential attacks. Existing studies [14] indicate that the gradient magnitude of the FC layer is related to the label count of the corresponding class. Such privacy information can be well protected by introducing the Gaussian mechanism into the LPI estimation. Moreover, each client only needs to compute and upload LPI once at the first round, followed by the standard FL communications. Therefore, each client with (ϵ, δ)-DP can effectively maintain a privacy-preserving LPI via a single query.

4. Experiment

4.1. Experimental Setup

Datasets. We conduct experiments on three long-tailed datasets: CIFAR-10-LT, CIFAR-100-LT [2], and ImageNet-LT [20]. Following [20], we report the top-1 accuracy of the global model on the balanced test set under different imbalance ratios (IR=50 and IR=100), as well as accuracy for many-shot, medium-shot, and few-shot classes.

FL Settings. We consider two popular settings to simulate heterogeneity [40]: (1) Dirichlet partition [9]. For each class in the global dataset, we first generate $\mathbf{p}_c \sim Dir(\alpha_{dir})$ for class c, then allocate p_c^k proportion of the samples in class c to client k. α_{dir} is a hyper-parameter for the degree of data heterogeneity. A smaller α_{dir} indicates a higher heterogeneity degree. We set α_{dir} as 0.5 for CIFAR-10-LT, 0.1 for CIFAR-100-LT, and 0.05 for ImageNet-LT. The visualizations of data distribution are shown in Section 9 in **Supplementary Material**. (2) Pathological partition [21, 41]. Each client is randomly assigned limited classes from all classes, *i.e.*, 3 out of 10 classes for CIFAR-10-LT and 10 out of 100 classes for CIFAR-100-LT.

Implementations. We adopt ResNet-18 [11] for CIFAR-10-LT, ResNet-32 [11] for CIFAR-100-LT, and ResNeXt-50 [38] for ImageNet-LT for fair comparisons. In each communication round, 20 clients are selected with 2 local epochs. More implementation details are reported in Section 6 in **Supplementary Material**.

Baselines. Including FedAvg [21] as the baseline, different methods are compared. (1) FL methods: FedProx [16], SCAFFOLD [15], FedAlign [22], and a client grouping method IFCA [8] for data heterogeneity, Ratio Loss[34] and CLIMB[30] for class imbalance; (2) centralized long-tailed methods: Focal Loss [19], CRT-IB, CRT-CB [14], LDAM [2], BSM [25], LADE [12], RIDE [35], and PaCo [4].

Table 1: Comparison results under Dirichlet partition.

Methods	CIFAR-10-LT		CIFAR-100-LT		ImageNet
	100	50	100	50	-LT
FedAvg [21]	53.16	61.67	34.48	36.84	33.80
FedProx [16]	61.43	71.78	34.16	38.77	32.98
SCAFFOLD [15]	58.11	69.96	34.94	37.07	34.23
FedAlign [23]	58.83	66.07	35.36	39.80	32.35
IFCA [8]	60.97	70.53	33.56	36.36	33.07
Ratio Loss [34]	53.31	62.26	33.06	34.94	33.15
CLIMB [30]	60.28	72.29	34.66	40.22	35.29
Focal Loss [19]	53.88	59.00	33.88	38.03	32.85
CRT-IB [14]	53.80	63.52	32.48	37.61	31.77
CRT-CB [14]	63.31	69.87	34.06	39.60	35.52
GBME	**71.07**	**76.85**	**40.42**	**45.16**	**45.75**

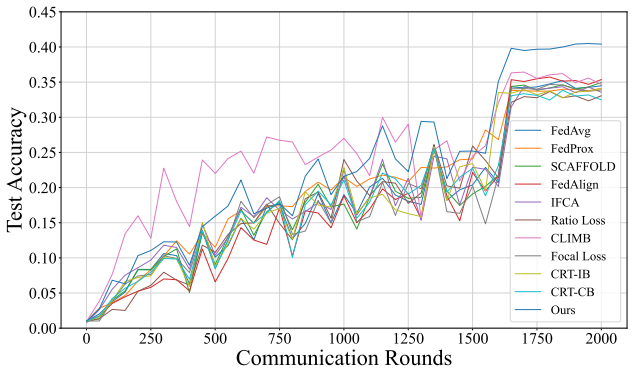

Figure 4: Accuracy tendency of GBME along with communication rounds on CIFAR-100-LT with IR = 100.

Table 2: Comparison results under pathological partition.

Methods	CIFAR-10-LT		CIFAR-100-LT	
	100	50	100	50
FedAvg [21]	52.79	56.58	33.90	37.73
FedProx [16]	53.55	56.08	33.07	38.04
SCAFFOLD [15]	54.68	54.24	34.14	39.00
FedAlign [23]	47.27	50.33	32.12	37.05
IFCA [8]	53.85	57.02	31.28	37.49
Ratio Loss [34]	49.94	51.01	34.54	38.48
CLIMB [30]	51.71	55.14	34.48	37.30
Focal Loss [19]	52.95	58.19	33.81	37.79
CRT-IB [14]	55.40	59.70	32.15	36.82
CRT-CB [14]	58.82	59.24	35.32	39.82
GBME	**63.07**	**60.19**	**38.84**	**43.46**

4.2. Main Results

Comparisons Results for Dirichlet Partition. The main results with the Dirichlet partition are summarized in Table 1. Overall, our approach GBME consistently outperforms

Table 3: Accuracy of different class priors, *i.e.*, GPI, local and global label distributions.

	Methods	CIFAR-10-LT		CIFAR-100-LT		ImageNet
		100	50	100	50	-LT
	FedAvg [21]	53.16	61.67	34.48	36.84	33.80
Local	LDAM [2]	63.55	69.50	32.54	37.01	30.70
Local	BSM [25]	66.64	74.29	35.55	39.50	34.39
Local	LADE [12]	66.05	74.54	33.28	39.89	36.32
Local	RIDE [35]	59.95	69.05	34.94	39.09	40.00
Local	PaCo [4]	65.11	71.35	35.92	40.03	38.26
Global	LDAM [2]	67.69	72.41	36.61	40.01	38.66
Global	BSM [25]	65.81	74.15	38.56	42.45	39.63
Global	LADE [12]	66.03	74.06	39.82	42.95	39.37
Global	RIDE [35]	60.37	68.39	34.59	39.35	40.81
Global	PaCo [4]	69.60	72.36	37.63	41.26	39.74
GPI	LDAM [2]	66.73	72.58	35.36	39.17	34.88
GPI	BSM [25]	67.44	74.36	37.19	41.91	37.64
GPI	LADE [12]	66.89	74.79	38.61	41.08	38.52
GPI	RIDE [35]	58.64	68.63	35.88	39.39	40.30
GPI	PaCo [4]	68.84	75.35	37.76	43.24	39.04
	GBME	**71.07**	**76.85**	**40.42**	**45.16**	**45.75**

Table 4: Accuracy of many/medium/few classes using local label distribution (Local), global label distribution (Global) and GPI on CIFAR-100-LT with IR = 100.

	Methods	Many	Medium	Few	Average
	FedAvg [21]	62.03	32.26	4.93	34.48
Local	LDAM [2]	56.23	29.91	7.97	32.54
Local	BSM [25]	57.63	36.34	8.87	35.55
Local	LADE [12]	54.00	33.74	8.57	33.28
Local	RIDE [35]	**64.97**	31.11	4.37	34.94
Local	PaCo [4]	35.60	**46.34**	24.13	35.92
Global	LDAM [2]	52.03	37.63	17.43	36.61
Global	BSM [25]	51.54	41.49	20.00	38.56
Global	LADE [12]	52.57	43.34	20.83	39.82
Global	RIDE [35]	63.29	32.34	3.73	34.59
Global	PaCo [4]	42.06	41.23	**28.27**	37.63
GPI	LDAM [2]	56.14	33.80	12.93	35.36
GPI	BSM [25]	52.80	39.26	19.07	37.19
GPI	LADE [12]	55.34	40.46	12.20	38.61
GPI	RIDE [35]	63.66	34.91	4.60	35.88
GPI	PaCo [4]	45.94	42.23	22.67	37.76
	GBME	49.97	44.14	24.93	**40.42**

previous works under different imbalance ratios. Compared with FedAvg, previous approaches obtain a small improvement on the large-scale datasets (CIFAR-100 and ImageNet), while our method can outperform previous solutions by a large margin. As shown in Figure 4, our method performs better than most baselines at any communication round and outperforms all comparison methods after about 1600 communication rounds, when the learning rate decays.

Comparisons Results for Pathological Partition. The pathological partition is a challenging setting for FL. As shown in Table 2, the proposed GBME still achieves competitive accuracy under such setting, which outperforms all comparison methods. More comparisons of our method are shown in Section 7 in **Supplementary Material**.

Table 5: Accuracy comparisons of GBME-p on CIFAR-100-LT (IR=100).

Methods	ϵ	Accuracy
LDAM [2]	-	35.36
BSM [25]	-	37.19
LDAE [12]	-	38.61
RIDE [35]	-	35.88
PaCo [4]	-	37.76
GBME-p	5	38.50
	10	39.27
	20	39.88
	100	40.24

Effectiveness of GPI. We compare the different priors including GPI, local and global label distributions. As shown in Table 3, using global label distribution usually performs better than using local label distribution, which is consistent with our theoretical conclusions (Lemma 1 and Theorem 1). Besides, using GPI exhibits competitive results with global label distribution (unknown due to privacy). Our method can obtain the best result on each dataset. Besides, we visualize the class-wise GPI curves on CIFAR-100-LT in Section 8 in **Supplementary Material**. GPI can improve the performance of minority classes and reduce the performance drop of majority classes, as it exhibits a similar yet flatter tendency compared with the global label distribution. Thus, GPI is effective for federated long-tailed learning.

Evaluation on Minority Classes. To better understand the improvement of our method, we report the accuracy for many/medium/few-shot classes on CIFAR-100-LT with IR = 100 in Table 4. Our method exhibits superior performance for medium-shot and few-shot classes. For re-balance methods (*i.e.*, LDAM, BSM and LADE), using GPI can improve the performance for medium-shot and few-shot classes compared with local label distribution. Similarly, using GPI can increase the accuracy for medium-shot and many-shot classes for RIDE and PaCo, while the improvement is slight for few-shot classes since RIDE and PaCo focus on representation quality over all classes rather than merely minority classes. Compared with global label distribution, previous re-balance methods with GPI can produce competitive results on minority classes.

Evaluation with Privacy Protection. In Table 5, we report the comparison results of GBME-p with various protection levels ϵ. The δ is fixed as 0.03. It is observed that

Table 6: Accuracy with the standard deviation.

Dataset	CIFAR-100-LT		CIFAR-10-LT		ImageNet-LT
IR	100	50	100	50	-
FedAvg [21]	34.02±0.52	36.95±0.46	55.76±3.64	65.09±3.18	33.96±0.56
SCAFFOLD [15]	34.58±0.45	37.69±0.54	60.33±2.15	70.99±0.93	34.21±0.38
GBME	40.14±0.54	44.66±0.47	73.12±1.86	77.24±1.48	46.02±0.33

Table 7: Comparisons under the IID distribution.

IID	CIFAR-100-LT		CIFAR-10-LT	
IR	100	50	100	50
FedAvg [21]	39.29	44.82	55.79	71.73
GBME	44.16	49.05	72.39	79.08

Table 8: Comparisons between GBME with 1000 communication rounds and FedAvg with 2000 rounds.

Method	CIFAR-100-LT		CIFAR-10-LT	
IR	100	50	100	50
FedAvg [21]	34.48	36.84	53.16	61.67
GBME	35.11	37.66	64.11	71.17

GBME-p with the Gaussian mechanism can still outperform previous solutions. Along with decreasing ϵ, the performance of GBME-p is also decreasing since we restrict the stronger privacy guarantees. As shown in Figure 6, GBME-p can keep lower similarity between the LPI and local label distribution, thus is able to protect the label privacy of the clients. Besides, GPI can work well as a global balanced prior because the similarity between GPI and global label distribution is very higher.

4.3. Further Analysis

Connection between Weight Norm and Gradient Summation in LPI. For a classifier trained on long-tailed data, the gradient magnitude for the weights of a class is correlated with the sample number of this class, *e.g.*, larger gradient summation of one update (orange line) on majority classes, resulting in larger weights and weight norms (blue line) on majority classes, as shown in Figure 5.

IID Data Distribution. In this section, we report the comparison results under the IID data distribution. Table 7 shows that the proposed GBME achieves much higher accuracy than FedAvg on IID data, telling that GBME can well handle long-tailed FL in both IID and non-IID cases.

Accuracy with the Standard Deviation. We conduct the experiments for multiple times and report the mean and the standard deviation. As indicated by Table 6, GBME can obtain the best accuracy compared with the baselines.

Computation Cost. Using FLOPs per round as the computation cost, with ResNet-32 on CIFAR-100-LT (IR=100), the basic cost of FedAvg is 1520.42G. Additional costs of

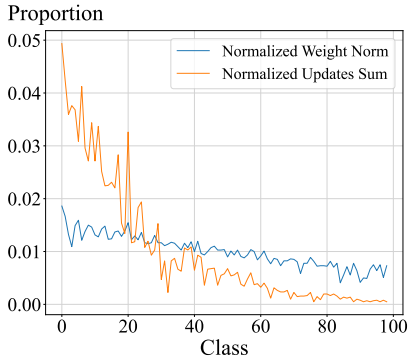

Figure 5: Curves of normalized weight norm and gradient sum on CIFAR-100-LT with IR=100.

GBME are from LPI calculation (*i.e.*, 760.21G) and multiple experts (*i.e.*, 696.06G for 3 experts). Since LPI is only computed in the first round, its cost is negligible but the accuracy gain is large (*i.e.*, 2.7%). The trade-off between the cost and benefit of multiple experts is controlled by the expert number. Adding one expert with 348G additional cost, the accuracy gain is about 1.6%.

Communication Cost. Using the parameters of a client as the communication cost, GBME (0.77M for 3 experts) has more parameters than FedAvg (0.46M). However, Table 8 shows that GBME can outperform FedAvg with only half communication rounds.

4.4. Ablation Studies

Component Analysis. Table 9 illustrates the component analysis of our method, including different class priors (GPI, local and global label distributions) as well as with and without the multi-expert architecture. It is observed that both class prior and multi-expert architecture can influence the final performance. In detail, taking global label distribution as the prior can obtain the best result, because it is ground-truth information of the dataset but unavailable in practice. Using GPI can achieve similar accuracy compared with global label distribution. Both of them outperform the local label distribution, which is consistent with our theoretical results. The performance can be further improved when combining GPI and the multi-expert architecture.

Different Client Grouping Strategies. In this part, we analyze the effects of different client grouping strategies, involving: randomly grouping clients (Random strategy), grouping clients by the cosine similarity between local and global label distributions (Label strategy), and grouping clients by the cosine similarity between LPI and GPI (GPI strategy). The main results are shown in Table 10, which indicate that the random strategy contributes little to the per-

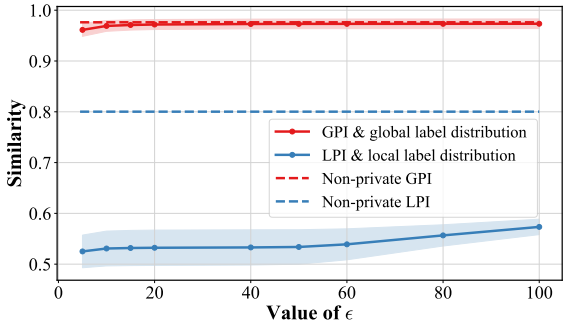

Figure 6: Visualizations of the similarity between the proxy information and label distribution for GBME-p.

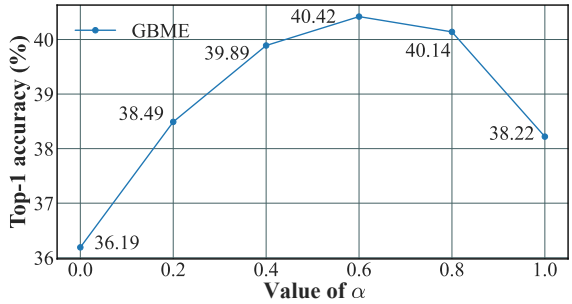

Figure 7: GBME with varying α values for client selection on CIFAR-100-LT with IR = 100.

Table 9: Component analysis for GBME, including different priors with and without multi-expert ensemble.

Methods	Prior	Ensemble	CIFAR-100-LT		ImageNet -LT
			100	50	
FedAvg [21]	✗	✗	34.48	36.84	33.80
BSM [25]	local	✗	35.55	39.50	34.39
BSM [25]	global	✗	38.56	42.45	39.63
BSM [25]	global	✓	**41.77**	**45.32**	45.57
BSM [25]	GPI	✗	37.19	41.91	37.64
BSM [25]	GPI	✓	40.42	45.16	**45.75**

Table 10: GBME with different grouping strategies and loss functions on CIFAR-100-LT with IR = 100.

Methods	Ensemble	Group strategy	Accuracy
LDAM [2]	✓	-	35.36
LDAM [2]	✓	GPI	38.25
BSM [25]	✓	-	37.19
BSM [25]	✓	Random	37.38
BSM [25]	✓	Label	40.51
BSM [25]	✓	GPI	40.42

formance, while the label and GPI strategy bring about a 3% improvement in accuracy, respectively. Thus, GPI for grouping clients can improve the final performance for re-balance strategies, such as BSM [25] and LDAM loss [2].

Table 11: GBME with different expert numbers on CIFAR-100-LT.

Expert Number	IR = 100	IR = 50	Parameters(Million)
$M = 1$	37.19	41.91	0.46
$M = 2$	38.71	44.92	0.52
$M = 3$	40.42	45.16	0.77
$M = 4$	41.68	46.88	1.02
$M = 5$	41.85	46.59	1.27

Varying α for Client Selection. As shown in Figure 7, the relationship between the hyper-parameter α and the accuracy exhibits a non-linear tendency. $0 < \alpha < 1$ can promote the information interaction among different groups via a multiple selection, where $\alpha \approx 0.6$ maximums such interaction power to obtain the higher performance. If $\alpha = 1$, each group only updates the corresponding expert and the client cannot interact with other groups. Hence it exhibits a lower accuracy due to heterogeneity. If $\alpha = 0$, the client randomly updates an expert at each round, which also performs a lower accuracy due to the limited interaction.

Different Expert Numbers. The effect of expert number M are shown in Table 11. It can be observed that a larger M results in the better performance but more parameters. Considering the trade-off between performance and communication cost, $M = 3$ and $M = 4$ are both good.

5. Conclusions

In this work, we propose a novel global balanced multi-expert (GBME) framework to address federated long-tailed problem. In particular, a proxy is designed as the class prior for existing re-balance algorithms to optimize a global balanced objective without requiring local label distributions. Such proxy can also guide the client grouping and selection to aggregate the heterogeneous knowledge in an ensemble manner. Moreover, we present a GBME-p algorithm with a theoretical guarantee, which equips the privacy protection ability with the concept of differential privacy. Extensive experiments on long-tailed decentralized datasets demonstrate the effectiveness of our method, both GBME and GBME-p showing superior performance to SOTA methods.

Acknowledgement Li Shen is supported by STI 2030—Major Projects (No. 2021ZD0201405). Li Liu was supported by the National Natural Science Foundation of China under grant No. 62101351, and the GuangDong Basic and Applied Basic Research Foundation under grant No. 2020A1515110376. Baoyuan Wu was supported by the National Natural Science Foundation of China under grant No. 62076213, Shenzhen Science and Technology Program under grants No. RCYX20210609103057050, No. ZDSYS20211021111415025, No. GXWD20201231105722002-20200901175001001.

References

[1] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–121, 2021. 3

[2] 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. 2, 6, 7, 9

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. 2

[4] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021. 2, 3, 6, 7

[5] 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. 2

[6] Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020. 1, 2

[7] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2014. 2, 5

[8] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020. 6

[9] Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, et al. Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*, 2020. 6

[10] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[12] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636, 2021. 2, 6, 7

[13] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, 2016. 1

[14] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019. 2, 3, 6

[15] 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. 2, 6, 8

[16] 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. 6

[17] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 133–141. Springer, 2019. 1

[18] Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2020. 2

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 6

[20] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 2, 6

[21] 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. 1, 2, 4, 5, 6, 7, 8, 9

[22] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022. 6

[23] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022. 6

[24] 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*, 2020. 5

[25] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33:4175–4186, 2020. 2, 5, 6, 7, 9

[26] Sumudu Samarakoon, Mehdi Bennis, Walid Saad, and Mérouane Debbah. Distributed federated learning for ultra-reliable low-latency vehicular communications. *IEEE Transactions on Communications*, 68(2):1146–1159, 2019. 1

[27] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-i.i.d. data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2020. 2

[28] Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022. 2

[29] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision*, pages 467–482. Springer, 2016. 2

[30] Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *International Conference on Learning Representations*, 2021. 1, 2, 6

[31] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[32] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007. 2

[33] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020. 2

[34] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10165–10173, 2021. 1, 2, 6

[35] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2020. 2, 3, 6, 7

[36] Penghui Wei, Hongjian Dou, Shaoguo Liu, Rongjun Tang, Li Liu, Liang Wang, and Bo Zheng. Fedads: A benchmark for privacy-preserving cvr estimation with vertical federated learning. *arXiv preprint arXiv:2305.08328*, 2023. 1

[37] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer, 2020. 3

[38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 6

[39] Miao Yang, Ximin Wang, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 2174–2178. IEEE, 2021. 1, 2

[40] 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. 6

[41] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2020. 6

[42] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems*, 35:34077–34090, 2022. 3

[43] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01):1–20, 2023. 2

[44] Yuxuan Zhang, Lei Liu, and Li Liu. Cuing without sharing: A federated cued speech recognition framework via mutual knowledge distillation. *arXiv preprint arXiv:2308.03432*, 2023. 1

[45] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 2

[46] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020. 3