Similarity-based Soft Supervision and Classifier Adjustment for Federated Long-Tailed Learning

Peng Wang, Liangcheng Wang, Hanlei Li, Jun Zhou*
Southwest University
Chongqing, China
{wp5066, wlc15284762136, paoxia333}@email.swu.edu.cn, zhouj@swu.edu.cn

***Abstract*—Federated learning usually assumes that global class distributions are balanced. However, real-world data exhibit a long-tail class distribution, which can significantly degrade the model performance on tail classes. Existing federated long-tailed learning methods do not fully utilize the semantic relationship between classes to enhance tail class learning and often suffer from representation bias caused by imbalanced distributions. To solve this issue, we propose a novel federated long-tailed learning framework called FedSSCA, which improves model performance on tail classes through an inter-class knowledge transfer mechanism and adaptive classifier adjustment. Specifically, we introduce a semantic association-based inter-class similarity soft supervision mechanism, which uses classifier weights as class prototypes to compute inter-class similarity and promote knowledge transfer among classes. Furthermore, we incorporate a dynamic classifier adjustment strategy based on estimated class distributions to optimize the feature representation space and mitigate representation bias . Extensive experiments on the CIFAR-10/100-LT and Tiny-ImageNet-LT long-tailed datasets demonstrate that FedSSCA achieves significant performance improvements compared to existing federated long-tailed learning methods.**

***Index Terms*—federated learning, long-tailed learning**

I. Introduction

Federated learning (FL), as an emerging machine learning paradigm, achieves collaborative model training across multiple parties while protecting user data privacy [1]. In the federated learning framework, participants (clients) train models locally using private data without sharing original data and only upload model parameters to a central server for aggregation [2], [3].

The primary challenge faced by federated learning is the non-independent and identically distributed (Non-IID) nature of client data, where there are significant differences in the distribution of data classes between different clients [4]. Currently, many methods attempt to alleviate this problem, such as personalized federated learning, knowledge distillation, and improved optimization strategies [5]. However, these methods often assume that the global distribution of data classes is balanced. In real-world federated learning scenarios, data often exhibit significant long-tailed characteristics, where a few head classes contain a large number of samples, whereas most tail classes have a very limited number of samples [6].

*Corresponding author: Jun Zhou (zhouj@swu.edu.cn).

This work was supported by the National Natural Science Foundation of China (No. 22274134), the Natural Science Foundation of Chongqing, China (CSTB2022NSCQ-MSX1519).

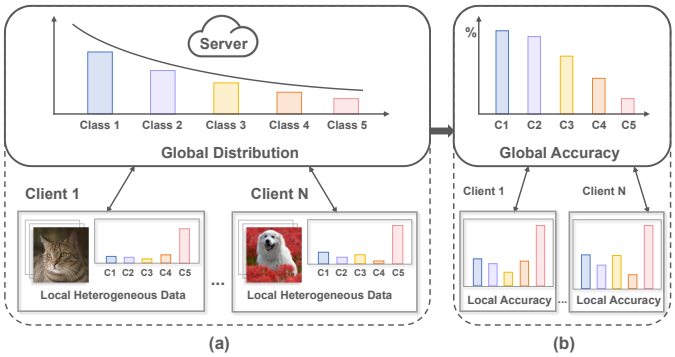

Fig. 1: Illustration of inconsistent global and local class distribution in federated long-tailed learning. (a) Tail classes may appear as head classes on certain clients. (b) Tail class identification accuracy globally and on clients.

In federated learning, long-tailed data distributions severely hinder the global model from effectively learning the features of tail classes, thereby reducing its ability to recognize these underrepresented categories [7]. Existing methods typically focus on re-balancing the training objective or correcting gradients but often treat classes independently. They underutilize inter-class semantic relationships, making it difficult to effectively transfer knowledge from semantic-rich head classes to data-scarce tail classes. Consequently, the learned feature representations for tail classes are often poor. Moreover, the privacy protection mechanism in federated learning limits the sharing of class distribution information, making it challenging to directly assess the degree of imbalance. This leads to severe classifier bias, where the decision boundaries of the aggregated global model are skewed towards head classes, suppressing the prediction probabilities of tail classes.

To address these challenges, we propose the FedSSCA framework, which enhances model performance on tail classes through an inter-class knowledge transfer mechanism and a post-training classifier adjustment strategy. The core idea of FedSSCA is to leverage semantic similarities to improve tail class representation and to rectify the classifier bias based on estimated distributions without violating privacy principles.

Our main contributions are as follows:

- We introduce a class-similarity soft supervision mechanism based on semantic associations. By using classi-

fier weights as class prototypes to measure inter-class similarities globally, this mechanism leverages inter-class relationships to facilitate knowledge transfer. This allows the model to learn more robust features for tail classes by borrowing semantic information from similar head classes.

- We propose a distribution-based classifier adjustment strategy. Leveraging the correlation between classifier weight norms and class sample sizes, we estimate the global class distribution without accessing raw data. Based on these estimates, we dynamically adjust the classifier weights to optimize the feature representation space, effectively mitigating the representation bias caused by long-tail distributions.
- Experiments on several standard long-tailed datasets (CIFAR-10/100-LT and Tiny-ImageNet-LT) demonstrate that FedSSCA significantly outperforms existing methods, achieving balanced performance improvements across all classes.

II. Related Work

A. Long-Tailed Learning

The long-tailed distribution is a common data imbalance phenomenon in the real world. Long-tailed learning methods are mainly categorized into three types [6]. Data-level methods, such as resampling [8], [9], balance training data by adjusting the sample sizes of different classes. Algorithm-level methods, such as re-weighting [10], [11], assign different weights to classes to emphasize tail classes. Loss function modification methods, including Focal Loss [12], Class-Balanced Loss [13], and LDAM [14], adjust the training objective to address imbalance. Recent studies [15], [16] propose representation learning strategies to mitigate class imbalance issues. However, these methods struggle to be directly applied in federated settings due to the lack of direct access to global data distribution information.

B. Federated Long-Tailed Learning

Federated learning enables collaborative model training among participants without sharing raw data. Existing methods typically assume a balanced global data distribution and focus on data heterogeneity across clients, failing to effectively handle the prevalent long-tailed distribution in real-world scenarios. Federated long-tailed learning aims to address the global data long-tailed distribution problem in federated environments. In recent years, several novel approaches have been proposed to tackle this challenge. RUCR [17] focuses on representation unification using global prototypes and achieves classifier correction through prototype mixing. Fed-GraB [18] introduces an adaptive gradient balancer that dynamically adjusts positive and negative gradient weights based on estimated global distribution, effectively alleviating class imbalance. GBME [15] leverages gradient-derived proxies as class priors and employs a multi-expert framework to aggregate heterogeneous knowledge from different clients. FedLoGe [19] trains models using the neural collapse principle to balance performance across classes.

III. Proposed Method

A. Preliminaries

Federated Learning. We consider a federated learning system with K clients, where $D = \bigcup_{k=1}^{K} D_k$ represents the union of all client data [1]. In each communication round, selected clients train locally on D_k and upload updated parameters to the server. The server aggregates these updates as: $w^{(t+1)} = \sum_{k \in S_t} \left(\frac{|D_k|}{|D|} \right) w_k^{(t)}$, where $w^{(t)}$ denotes global model parameters at round t, and S_t is the set of participating clients.

Long-tailed Distribution. In long-tailed scenarios, sample counts per class show significant imbalance. Let n_c denote samples in class c. Classes arranged by descending sample count satisfy $n_1 \geq n_2 \geq \ldots \geq n_C$, where C is the total number of classes. The Imbalance Factor (IF) [14] measures this imbalance: $\text{IF} = \frac{n_{\max}}{n_{\min}}$, where $n_{\max}$ and $n_{\min}$ are the sample counts of the most and least frequent classes.

B. Overview of FedSSCA

To address the insufficient recognition capability for tail classes in federated learning caused by skewed label distributions and global data imbalance, we propose the FedSSCA framework. As illustrated in Fig 2, the core idea of FedSSCA is to enhance the model's performance through two synergistic components: an Inter-class Similarity-guided Soft Labeling (SSL) mechanism during local training, and a Distribution-based Classifier Adjustment (DCA) strategy after global aggregation.

C. Norm-based Distribution Estimation

In federated learning environments, due to privacy protection mechanisms, servers cannot directly access the original data from clients or their actual class distribution information. This limitation makes it challenging to apply traditional long-tailed learning methods that rely on global class frequencies. However, recent research has discovered a significant positive correlation between the magnitude of the weight norm of each class in the classifier and the number of samples in that class [20], [21]. This phenomenon provides an effective approach to indirectly infer class distribution imbalance without leaking raw data.

Leveraging this finding, we employ a norm-based strategy to estimate the global class distribution, which serves as the basis for our subsequent classifier adjustment. Specifically, we first calculate the L_2 norm of the weight vector for each class in the global classifier:

$$\|W_c\|_2 = \sqrt{\sum_{j=1}^{d} (W_{c,j})^2} \tag{1}$$

where $||W_c||_2$ represents the L_2 norm of the classifier weight vector corresponding to class c, and d is the dimension of the weight vector.

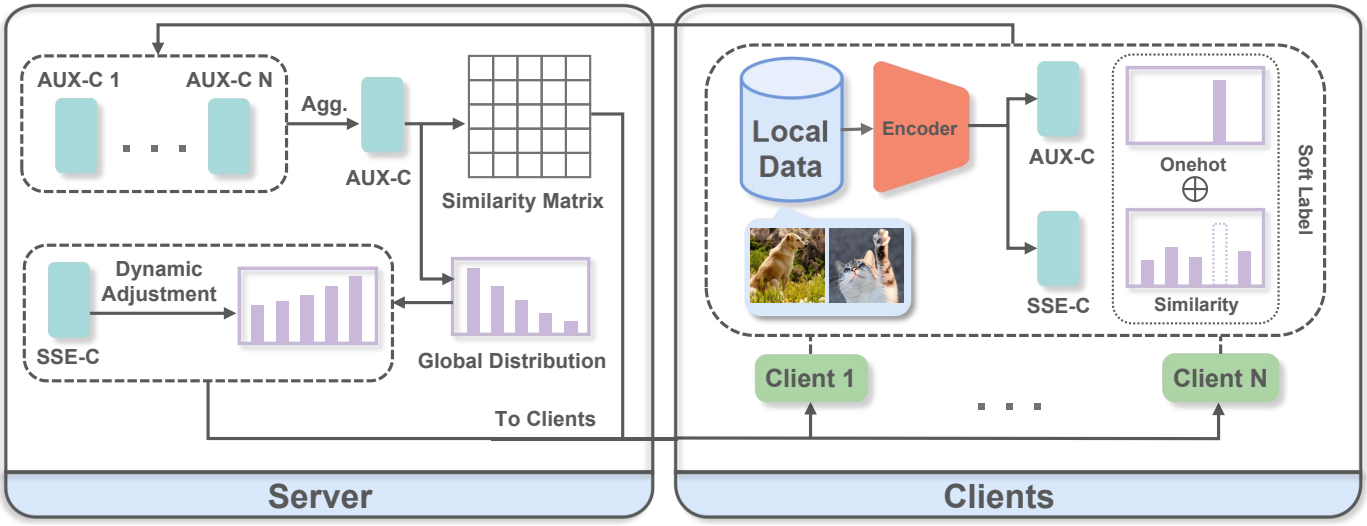

Fig. 2: Overview of the FedSSCA Framework. Our method integrates two core components: inter-class similarity soft labeling and distribution-based classifier adjustment. Here, the global classifier (AUX-C) is utilized for global distribution estimation to guide the dynamic adjustment. SSE-C serves as an equiangular tight frame classifier for feature optimization.

To reduce the impact of norm scale variations across different training stages, we apply mean normalization to the norms:

$$\tilde{W}_c = \frac{\|W_c\|_2}{\frac{1}{C}\sum_{i=1}^{C}\|W_i\|_2} \tag{2}$$

where C is the total number of classes, and $\tilde{W}_c$ is the normalized norm value. Finally, we convert these normalized norms into a probability distribution estimation q_c^g through a softmax function:

$$q_c^g = \text{softmax}(\tilde{W}_c) = \frac{\exp(\tilde{W}_c)}{\sum_{i=1}^{C}\exp(\tilde{W}_i)} \tag{3}$$

this estimated global distribution q_c^g effectively captures the imbalance characteristics of the data and provides the necessary guidance for the DCA introduced in Section III-E.

D. Inter-class Similarity-guided Soft Labeling

To effectively capture semantic relationships between classes and mitigate overfitting to majority classes in hard-label training, we introduce a Similarity-guided Soft Labeling mechanism that provides soft supervision by leveraging inter-class semantic similarities.

Prototype-based Similarity Computation. In the federated setting, accessing raw data violates privacy principles. We treat the classifier's weight vectors as class prototypes $P = [p_1, ..., p_C] \in \mathbb{R}^{C\times d}$, where p_i represents the prototype vector of the i-th class. We quantify semantic relationships by computing cosine similarity between these prototype vectors, yielding a similarity matrix $S \in \mathbb{R}^{C\times C}$: $S = PP^T$, where $S_{i,j}$ represents the semantic similarity between class i and class j.

Soft Label Generation. Using the computed similarity matrix, we transform one-hot encoded hard labels into soft labels that incorporate semantic relationships. For a sample with ground-truth label y, its soft label $\tilde{y}$ is:

$$\tilde{y} = (1-\beta)\cdot y_{\text{onehot}} + \beta \cdot \frac{e^{S_y}}{\sum_j e^{S_{yj}}} \tag{4}$$

where $\beta \in [0, 1]$ controls the degree of softening. This ensures the label distribution emphasizes the target class while assigning non-zero probabilities to semantically similar classes.

Knowledge Transfer Mechanism. The introduction of soft labels alters the gradient flow during backpropagation. With soft label loss, the gradient becomes:

$$\frac{\partial L_{\text{soft}}}{\partial w_i} = \left[(1-\beta)(\hat{y}_i - y_{\text{onehot},i}) + \beta\left(\hat{y}_i - \frac{e^{S_{i,j}}}{\sum_k e^{S_{k,j}}}\right)\right]\cdot f \tag{5}$$

this mechanism shows that weight updates now depend not only on prediction error but also on semantic similarity to the ground-truth class. For tail classes sharing semantic features with head classes, this facilitates knowledge transfer, enabling the model to learn richer representations even with limited samples.

E. Distribution-based Classifier Adjustment

To mitigate representation bias where the model overfits head classes while suppressing tail classes, we propose a Distribution-based Classifier Adjustment mechanism. While SSL enhances feature learning during local training, the classifier weights may still exhibit significant bias due to imbalanced global distribution. DCA addresses this by adaptively optimizing the L_2 norm of weight vectors based on the global distribution estimated in Section III-C.

Adjustment Factor Construction. We construct class-specific adjustment factors derived from the estimated global

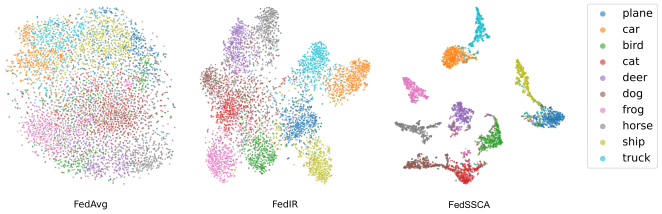

Fig. 3: t-SNE visualization of feature embeddings on CIFAR10-LT test set comparing our method with baseline approaches.

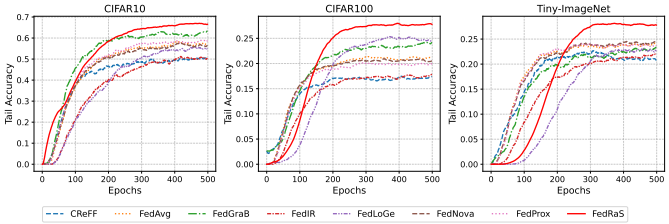

Fig. 4: Comparison of tail-class accuracy on CIFAR10-LT, CIFAR100-LT, and Tiny-ImageNet-LT.

probability distribution q_c^g. The adjustment factor A_c for class c is:

$$A_c = \gamma \cdot \left(\log \left(\frac{1}{q_c^g} \right) \right)^{\theta} \tag{6}$$

where γ controls the overall scale and θ controls the degree of non-linearity. For tail classes with smaller q_c^g, A_c becomes larger to enhance their weight norms, while head classes receive smaller A_c to constrain their influence.

Dynamic Weight Recalibration. Based on the adjustment factors, we recalibrate the weight vector w_c for each class c:

$$w_c' = w_c \cdot \frac{A_c}{||w_c||_2} \tag{7}$$

this operation forces the weight magnitude to align with the class-dependent target A_c, achieving information enhancement based on class scarcity. Unlike methods using fixed scaling factors, our approach leverages dynamically estimated global distribution, enabling refined and adaptive regulation that evolves as the global model converges, continuously optimizing decision boundaries to balance performance between head and tail classes.

IV. Experiments

A. Experimental Setup

We evaluate our method on three long-tailed benchmarks with varying imbalance factors (IF): CIFAR10-LT and CIFAR100-LT [14], [22] with IF $\in \{10, 50, 100\}$, and Tiny-ImageNet-LT [23] (200 classes, 64×64 images) with IF=10.

For non-IID simulation, we use Dirichlet distribution partitioning with $\alpha \in \{0.5, 1.0\}$ for CIFAR10-LT and $\alpha = 0.5$ for CIFAR100-LT and Tiny-ImageNet-LT, where lower α indicates stronger heterogeneity. We deploy 40 clients for CIFAR10-LT and 20 for others, with 5 local training rounds before aggregation. Network architectures are ResNet18 [24] for CIFAR10-LT and ResNet34 for CIFAR100-LT and Tiny-ImageNet-LT.

We compare against classic federated learning baselines (FedAvg [1], FedProx [25], FedNova [26]) and specialized federated long-tailed methods (FedIR [27], CReFF [7], FedGraB [18], FedLoGe [19]).

B. Experimental Results and Analysis

Table I demonstrates that FedSSCA consistently outperforms all baselines on CIFAR10-LT. Under the most challenging setting (IF= 100, $\alpha = 0.5$), our method achieves $\sim 1.3\%$ higher accuracy than the second-best approach. Table II extends this to CIFAR100-LT and Tiny-ImageNet-LT, where FedSSCA maintains superiority with up to 4.3% improvement on Tiny-ImageNet-LT.

Fig. 4 reveals a significant performance gap in tail-class recognition, particularly pronounced on Tiny-ImageNet-LT, validating that our soft supervision mechanism effectively transfers knowledge to data-scarce categories. Training curves show a notable performance jump around the 100th communication round, coinciding with the stabilization of our Norm-based Distribution Estimation. As the estimated distribution converges to the true distribution, the DCA mechanism precisely recalibrates classifier weights, correcting decision boundaries previously biased toward head classes. This enables the model to recover tail-class performance without sacrificing head-class accuracy—a capability baseline methods lack.

C. Visualization Analysis

Fig. 3 presents t-SNE visualization of CIFAR10-LT test set feature embeddings. Compared with FedAvg and FedIR, FedSSCA achieves significantly more distinct class separation and learns compact, discriminative representations for tail classes. This improvement is attributed to our SSL mechanism, which introduces soft targets based on semantic prototypes during local training, preventing overfitting to noisy or scarce labels.

Fig. 5 (Left) demonstrates high alignment between the true global distribution and our norm-based estimation, particularly for tail classes. This accuracy is critical as DCA relies on these estimated probabilities (p_c^g) to compute norm adjustment factors. The right plot shows the CDF comparison, where curves intersect near class 3, indicating that without adjustment, the model overestimates probabilities for head classes while underestimating them for middle and tail classes. This validates the necessity of DCA—by mapping these deviations to adjustment factors (A_c), FedSSCA effectively counteracts the observed bias, providing a solid foundation for classifier rectification.

TABLE I: Performance comparison with detailed class-wise results. Accuracy comparison between FedSSCA and baseline methods under different α and IF settings across Many-shot, Medium-shot, Few-shot, and All classes.

Setting	Method	IF=10				IF=50				IF=100			
		Many	Med	Few	All	Many	Med	Few	All	Many	Med	Few	All
α=0.5	FedAvg	0.8764	0.8635	0.8290	0.8665	0.8570	0.7597	0.6963	0.7796	0.9067	0.7007	0.5577	0.7053
	FedProx	0.8732	0.8650	0.8720	0.8698	0.8527	0.7800	0.6607	0.7733	0.8930	0.7337	0.5755	0.7182
	FedNova	0.8734	0.8760	0.8520	0.8723	0.8477	0.7930	0.6940	0.7852	0.8973	0.6710	0.6632	0.7358
	FedIR	0.8546	0.8552	0.8530	0.8547	0.8392	0.7090	0.6400	0.7404	0.8697	0.6467	0.4882	0.6502
	CReFF	0.8630	0.8585	0.9370	0.8686	0.8615	0.7340	0.7163	0.7797	0.8887	0.7037	0.6187	0.7252
	Fed-GraB	0.8676	0.8862	0.8530	0.8736	0.8652	0.8047	0.7273	0.8057	0.8793	0.7260	0.6587	0.7451
	FedLoGe	0.8796	0.8497	0.8940	0.8691	0.8430	0.8253	0.7583	0.8123	0.8873	0.7047	0.6715	0.7462
	Ours	0.8846	0.8805	0.9380	**0.8883**	0.8625	0.8180	0.7613	**0.8188**	0.8897	0.7257	0.6853	**0.7587**
α=1	FedAvg	0.8632	0.8735	0.8640	0.8674	0.8677	0.7850	0.6817	0.7871	0.9167	0.7007	0.6250	0.7352
	FedProx	0.8824	0.8655	0.8710	0.8745	0.8762	0.7950	0.6813	0.7934	0.9203	0.7233	0.6200	0.7411
	FedNova	0.8948	0.8540	0.8540	0.8744	0.8640	0.8187	0.7123	0.8049	0.9127	0.7303	0.6195	0.7407
	FedIR	0.8874	0.8410	0.8380	0.8639	0.8555	0.7287	0.6740	0.7630	0.9083	0.6700	0.5570	0.6963
	CReFF	0.8862	0.8535	0.8640	0.8709	0.8522	0.7760	0.7063	0.7856	0.9030	0.6700	0.5807	0.7042
	Fed-GraB	0.9070	0.8522	0.8900	0.8834	0.8900	0.7980	0.7257	0.8131	0.9153	0.7537	0.6530	0.7619
	FedLoGe	0.8882	0.8812	0.8980	0.8864	0.8870	0.8120	0.7150	0.8129	0.9103	0.7133	0.6685	0.7545
	Ours	0.9046	0.8755	0.9250	**0.8950**	0.8925	0.8213	0.7477	**0.8277**	0.9030	0.7590	0.6903	**0.7747**

TABLE II: Top-1 classification accuracy comparison on CIFAR100-LT and Tiny-ImageNet-LT under different settings.

Method	CIFAR100-LT								Tiny-ImageNet-LT			
	IF=50				IF=100				IF=10			
	Many	Med	Few	All	Many	Med	Few	All	Many	Med	Few	All
FedAvg	0.6550	0.4727	0.2300	0.4388	0.6085	0.4269	0.1853	0.3624	0.5400	0.3903	0.2569	0.4324
FedProx	0.6262	0.4357	0.2058	0.4093	0.6337	0.4454	0.1926	0.3774	0.5341	0.3963	0.2421	0.4289
FedNova	0.6412	0.4630	0.2145	0.4256	0.6393	0.4188	0.1972	0.3742	0.5240	0.3991	0.2656	0.4299
FedIR	0.5697	0.3847	0.1916	0.3705	0.5411	0.3585	0.1587	0.3139	0.4833	0.3491	0.2364	0.3882
CReFF	0.5997	0.4207	0.1837	0.3879	0.5963	0.4177	0.1694	0.3530	0.5314	0.3549	0.2421	0.4132
Fed-GraB	0.6522	0.4857	0.2516	0.4500	0.6548	0.4546	0.2072	0.3924	0.5187	0.3877	0.2538	0.4212
FedLoGe	0.6225	0.5163	0.2963	0.4667	0.6252	0.5212	0.2730	0.4326	0.5007	0.3783	0.2774	0.4143
Ours	0.6941	0.5290	0.2784	**0.4866**	0.7033	0.5177	0.2449	**0.4396**	0.5541	0.4240	0.2892	**0.4569**

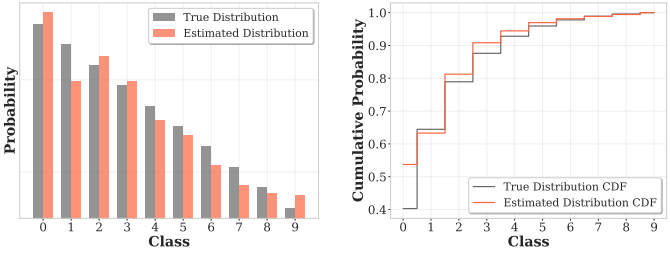

Fig. 5: Distribution analysis results showing comparison between true global distribution and estimated global distribution using our proposed method (left), and cumulative distribution functions comparison between estimated and true distributions (right).

TABLE III: Ablation study of different components in our method.

Components		Performance			
SSL	DCA	Many	Med	Few	All
–	–	0.9067	0.7007	0.5577	0.7053
✓	–	0.9107	0.7367	0.5677	0.7213
–	✓	0.8863	0.7490	0.6525	0.7516
✓	✓	0.8897	0.7257	0.6853	**0.7587**

TABLE IV: Performance comparison with different β values.

β	Many	Med	Few	All
0.1	0.8643	0.7203	0.6822	0.7483
0.2 (default)	0.8897	0.7257	0.6853	**0.7587**
0.3	0.8597	0.7190	0.6525	0.7346

D. Ablation Study

To verify the effectiveness of each component, we conducted ablation experiments on CIFAR10-LT (Table III). The SSL mechanism alone improves overall accuracy from 70.53% to 72.13%, with consistent gains on Medium and Few classes, confirming that incorporating semantic information from global prototypes helps learn more generalizable feature representations. The DCA strategy yields substantial improvements in tail-class accuracy by recalibrating weight norms based on estimated global distribution, rectifying classifier bias toward head classes. The complete FedSSCA framework achieves 75.87% overall accuracy with clear synergy: SSL optimizes the feature space while DCA aligns classifier boundaries. This combination delivers a remarkable 12.76% improvement in tail-class accuracy compared to baseline.

We evaluate the impact of key hyperparameters on CIFAR10-LT (IF= 100, $\alpha = 0.5$). As shown in Tables IV, V, and VI, our method demonstrates robust performance across

reasonable parameter ranges. For the soft labeling parameter β, the default value of 0.2 achieves the best balance—smaller values (0.1) provide insufficient knowledge transfer, while larger values (0.3) introduce noise from dissimilar classes. For classifier adjustment parameters θ and γ in Equation 6, $\theta = 1.4$ and $\gamma = 2.0$ offer optimal performance. Gentler adjustments ($\theta = 1.0$ or $\gamma = 1.0$) under-correct the tail-class bias, whereas aggressive adjustments ($\theta = 1.6$ or $\gamma = 3.0$) over-boost tail classes at the expense of head-class accuracy. These results confirm that our default hyperparameter settings provide an effective trade-off between head and tail class performance.

TABLE V: Performance comparison with different θ values.

θ	Many	Med	Few	All
1	0.8887	0.7473	0.6262	0.7413
1.2	0.8410	0.7570	0.6602	0.7435
1.4 (default)	0.8897	0.7257	0.6853	0.7587
1.6	0.8200	0.7310	0.7410	**0.7617**

TABLE VI: Performance comparison with different γ values.

γ	Many	Med	Few	All
1	0.8263	0.7343	0.6925	0.7452
1.5	0.8430	0.7177	0.7115	0.7528
2 (default)	0.8897	0.7257	0.6853	**0.7587**
2.5	0.8523	0.7210	0.7145	0.7578
3	0.8350	0.7363	0.6970	0.7502

V. Conclusion

This paper proposes a novel federated long-tailed learning framework called FedSSCA. To address the challenge of learning from imbalanced data, our approach integrates an inter-class knowledge transfer mechanism with an adaptive classifier adjustment strategy. Specifically, we introduce a similarity-guided soft labeling mechanism that leverages inter-class semantic relationships to enhance the model's generalization ability for tail classes. Furthermore, we optimize the feature space structure through a distribution-based classifier adjustment technique, which estimates global class distributions from classifier norms and adaptively recalibrates the classifier weights to mitigate representation bias. Extensive experiments demonstrate that FedSSCA achieves significant performance improvements on multiple standard long-tailed datasets.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 54, Apr. 2017, pp. 1273–1282.

[2] G. Long, T. Shen, Y. Tan, L. Gerrard, A. Clarke, and J. Jiang, *Federated Learning for Privacy-Preserving Open Innovation Future on Digital Health*. Springer International Publishing, 2022, pp. 113–133.

[3] C. Ren, H. Yu, H. Peng, X. Tang, B. Zhao, L. Yi, A. Z. Tan, Y. Gao, A. Li, X. Li, Z. Li, and Q. Yang, "Advances and open challenges in federated foundation models," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 2087–2126, 2026.

[4] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.

[5] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous federated learning: State-of-the-art and research challenges," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 79:1–79:44, Oct. 2023.

[6] C. Zhang, G. Almpanidis, G. Fan, B. Deng, Y. Zhang, J. Liu, A. Kamel, P. Soda, and J. Gama, "A systematic review on long-tailed learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 8, pp. 13 670–13 690, 2025.

[7] X. Shang, Y. Lu, G. Huang, and H. Wang, "Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022, pp. 2218–2224.

[8] J.-X. Shi, T. Wei, Y. Xiang, and Y.-F. Li, "How re-sampling helps for long-tail learning?" in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 75 669–75 687.

[9] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2021, pp. 5212–5221.

[10] X. Chen, Y. Zhou, D. Wu, C. Yang, B. Li, Q. Hu, and W. Wang, "Area: Adaptive reweighting via effective area for long-tailed classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2023, pp. 19 277–19 287.

[11] M. Li, Y.-m. Cheung, and Y. Lu, "Long-tailed visual recognition via gaussian clouded logit adjustment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 6929–6938.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct 2017.

[13] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2019.

[14] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Adv. Neural Inf. Process. Syst.*, 2019.

[15] Y. Zeng, L. Liu, L. Liu, L. Shen, S. Liu, and B. Wu, "Global balanced experts for federated long-tailed learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2023, pp. 4815–4825.

[16] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[17] W. Huang, Y. Liu, M. Ye, J. Chen, and B. Du, "Federated learning with long-tailed data via representation unification and classifier rectification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 5738–5750, 2024.

[18] Z. Xiao, Z. Chen, S. Liu, H. Wang, Y. Feng, J. Hao, J. T. Zhou, J. Wu, H. Yang, and Z. Liu, "Fed-GraB: Federated long-tailed learning with self-adjusting gradient balancer," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 77 745–77 757.

[19] Z. Xiao, Z. Chen, L. Liu, Y. FENG, J. T. Zhou, J. Wu, W. Liu, H. H. Yang, and Z. Liu, "Fedloge: Joint local and generic federated learning under long-tailed data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[20] D. Kim, K. Wang, S. Sclaroff, and K. Saenko, "A broad study of pre-training for domain generalization and adaptation," in *European Conference on Computer Vision*, 2022, pp. 621–638.

[21] C. Thrampoulidis, G. R. Kini, V. Vakilian, and T. Behnia, "Imbalance trouble: Revisiting neural-collapse geometry," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27 225–27 238.

[22] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[23] P. Chrabaszcz, I. Loshchilov, and F. Hutter, "A downsampled variant of ImageNet as an alternative to the CIFAR datasets," 2017, arXiv:1707.08819.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016.

[25] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.

[26] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7611–7623.

[27] T.-M. H. Hsu, H. Qi, and M. Brown, "Federated visual classification with real-world data distribution," in *European Conference on Computer Vision*, 2020, pp. 76–92.