

You Are Your Own Best Teacher: Achieving Centralized-level Performance in Federated Learning under Heterogeneous and Long-tailed Data

Shanshan Yan¹ Zexi Li^{3,4} Chao Wu⁴ Meng Pang⁵ Yang Lu^{1,2*} Yan Yan^{1,2} Hanzi Wang^{1,2}

¹Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, Xiamen, China

²Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China

³University of Cambridge ⁴Zhejiang University

⁵School of Mathematics and Computer Sciences, Nanchang University, Nanchang

yanshan@stu.xmu.edu.cn, zexi.li@zju.edu.cn, chao.wu@zju.edu.cn,
mengpang@ncu.edu.cn, luyang@xmu.edu.cn, yanyan@xmu.edu.cn, hanzi_wang@163.com

Abstract

Data heterogeneity, stemming from local non-IID data and global long-tailed distributions, is a major challenge in federated learning (FL), leading to significant performance gaps compared to centralized learning. Previous research found that poor representations and biased classifiers are the main problems and proposed neural-collapse-inspired synthetic simplex ETF to help representations be closer to neural collapse optima. However, we find that the neural-collapse-inspired methods are not strong enough to reach neural collapse and still have huge gaps to centralized training. In this paper, we rethink this issue from a self-bootstrap perspective and propose FedYoYo (You Are Your Own Best Teacher), introducing Augmented Self-bootstrap Distillation (ASD) to improve representation learning by distilling knowledge between weakly and strongly augmented local samples, without needing extra datasets or models. We further introduce Distribution-aware Logit Adjustment (DLA) to balance the self-bootstrap process and correct biased feature representations. FedYoYo nearly eliminates the performance gap, achieving centralized-level performance even under mixed heterogeneity. It enhances local representation learning, reducing model drift and improving convergence, with feature prototypes closer to neural collapse optimality. Extensive experiments show FedYoYo achieves state-of-the-art results, even surpassing centralized logit adjustment methods by 5.4% under global long-tailed settings. The code is available at <https://github.com/shanss132/FedYoYo>.

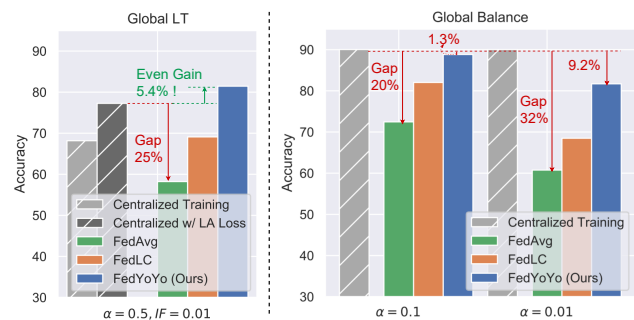


Figure 1. **Our method substantially closes the gap between centralized training and federated learning under heterogeneous data.** **Left:** non-IID data with global long-tailed (LT) distribution, where non-IID $\alpha = 0.5$ and imbalance factor (IF) = 0.01. **Right:** non-IID data with global balanced distribution (vanilla non-IID setting in the literature: the smaller α , the more non-IID). Baselines: vanilla centralized training, LA loss [21], FedAvg [20], FedLC [35].

1. Introduction

Federated learning (FL) [11, 14, 15, 20] is a collaborative learning paradigm that builds machine learning models from distributed data sources without sharing the raw data, which is communication-efficient [20] and privacy-preserving [34]. FL is promising in a wide range of scenarios, like medical imaging [1], multi-media analysis [30], the Internet of Things [39], etc. One inherent and key challenge in FL is data heterogeneity, a critical bottleneck that significantly impacts the performance of FL methods [14, 16].

In practice, the overall data distribution in an FL system is often long-tailed, meaning that data heterogeneity arises from both the local non-IID data and the global long-

*Corresponding Author: Yang Lu (luyang@xmu.edu.cn)

tailed [23, 24, 40] data. This combination leads to a more severe form of heterogeneity, causing in sub-optimal local models and poor generalization of the aggregated global model. As a result, there is always a huge gap between federated methods and centralized training due to data heterogeneity, which becomes even more evident when clients’ data are highly non-IID [11, 14, 16, 20].

Previous works tried to tackle non-IID data and improve the performances of FL by a large margin from FedAvg, but the centralized-federated gap is also dominant (Fig. 1). Recent studies [16, 17, 19, 28, 31] have shown that data heterogeneity can lead to poor feature representations and biased classifiers, which might be the main cause of bad generalization performances. The poor representations in local models further exacerbates the feature misalignment between the global model and other client models. Therefore, the key challenge is to learn better and more unified feature representations across clients with non-IID data distributions. Existing approaches, such as FedETF [16] and FedLoGe [31], attempt to mitigate the effects of heterogeneity by leveraging the theory of Neural Collapse [9, 12] and constructing a synthetic simplex equiangular tight frame (ETF) with maximal pairwise angles [22]. Neural collapse is a deep learning phenomenon, which depicts an ideal representation and classifier structure under balanced and sufficient training. Though these methods are inspired by neural collapse, we find they are not good enough to reach neural collapse optima under severe heterogeneity and global long-tailed distribution. An intuitive visualization is in Fig. 2: the pair-wise prototype angles of FedETF and FedLoGe are far away from the theoretical neural collapse optima.

Therefore, rethinking the representation learning strategy in non-IID federated learning is needed. We notice that self-supervised learning (SSL) has been widely validated in large-scale representation learning tasks, demonstrating the ability to capture more robust feature representations. This suggests that we can think of *supervised* FL in an *unsupervised* way to gain better representations, but more tailored designs are needed. The question is *how to effectively use class/label distributions under data heterogeneity*. Existing SSL uses sample-wise contrastive or bootstrap methods and learns feature extractors instead of the whole model. But in supervised heterogeneous FL, class-biased local data and long-tailed global data will make features prioritize majority classes while neglecting minority ones, also the local model representations are not aligned, causing severe model drifts.

In this paper, we find that self-bootstrap representation learning can release its full potential under the supervision of distribution-aware logit adjustment, and propose **FedYoYo: You Are Your Own Best Teacher** for local clients. Unlike previous SSL, we learn logits as representations instead of features, and logit adjustment can serve as distribution

guidance to improve the representation of minor classes and align the representations across clients. FedYoYo consists of two core components: Augmented Self-bootstrap Distillation (ASD) and Distribution-aware Logit Adjustment (DLA). ASD is inspired by BYOL [7] in SSL, but tailored designs are made in FL. We learn logits instead of features and use one local model as the teacher itself instead of two online and target models in BYOL. We use the logits of a weakly augmented sample as the teacher to guide the learning of a strongly augmented sample. DLA is a tailored version of logit adjustment for heterogeneous FL; considering the potential long-tailed global distribution, we realize a tradeoff between both local and global distributions.

FedYoYo realizes a matched co-design of self-bootstrap learning and logit adjustment, reaching near-centralized performances under non-IID FL (Fig. 1). In Fig. 2, it can be seen that our FedYoYo has better representations closer to neural collapse optimality and reaches higher generalization, compared with ETF-based methods.

Our main contributions are summarized as follows:

- We propose FedYoYo, which addresses two challenges: data heterogeneity and the combined issue of global long-tailed and local non-IID data. It has two key components: Augmented Self-bootstrap Distillation and Distribution-Aware Logit Adjustment.
- FedYoYo achieves a new level of performance in FL, it is comparable to centralized training. FedYoYo closes the centralized-federated gap from 20% to 1.3% in vanilla $\alpha = 0.1$ data heterogeneity, and it even surpasses the centralized method by 5.4% under global long-tailed distribution.
- We provide a new perspective on solving non-IID data and the caused poor representation issues in FL. We make FL more applicable and promising by reaching performances comparable to those of centralized.

2. Proposed Method

In this section, we introduce our proposed method FedYoYo, which focuses on enhancing feature representations and mitigating feature misalignment under data heterogeneity. The method consists of two key components: Augmented Self-bootstrap Distillation (ASD) and Distribution-aware Logit Adjustment (DLA). ASD employs a self-bootstrap mechanism with logit adjustment as distribution guidance to enhance feature extraction, while DLA calibrates the classifier outputs using a fused global-local distribution to address client bias. Together, these components enable consistent and robust feature representation across clients, as illustrated in the overall framework Fig. 3.

2.1. Preliminaries

In the context of FL, consider a scenario with K clients, where each client holds a non-IID local dataset $\mathcal{D}_k =$

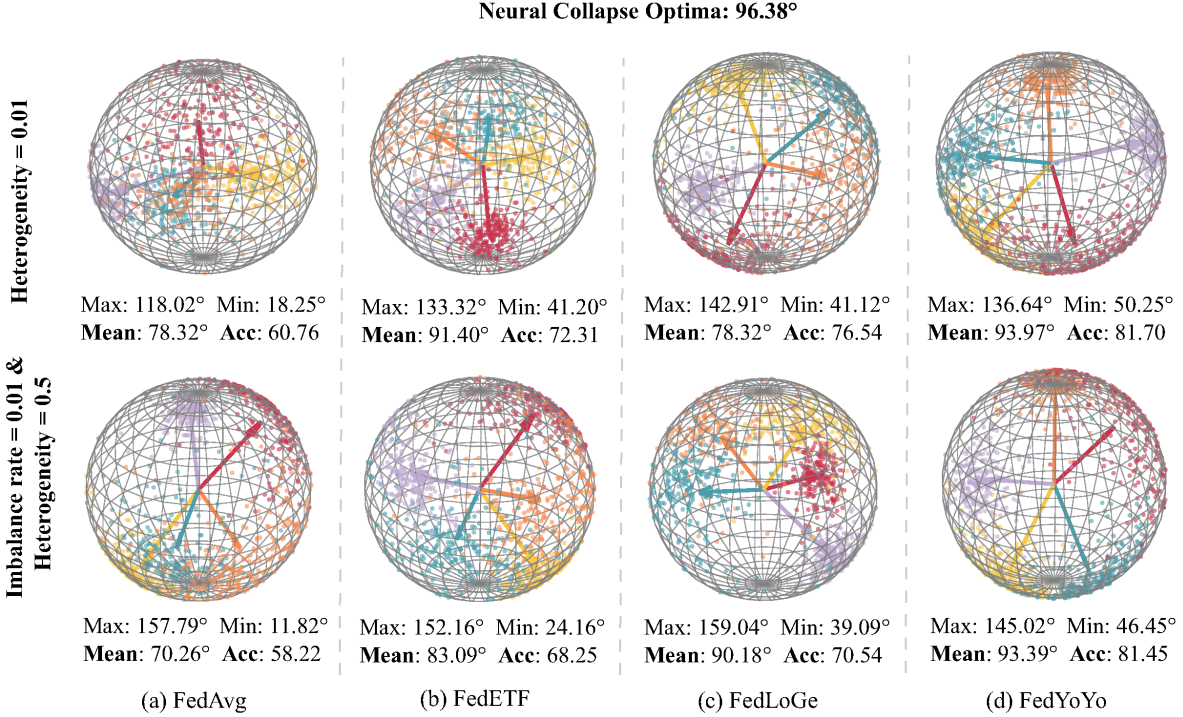


Figure 2. **Visualization of neural collapse degrees and accuracy for global models on CIFAR-10.** The optimal collapse angle (96.38°) follows Equiangular Tight Frame (ETF) theory, which defines ideal class prototype angles for maximal separation. Arrows indicate prototype representations, and colors denote different categories. “Max”, “Min”, and “Mean” represent the largest, smallest, and average angles between class prototypes, while “Acc” denotes global model accuracy. The top row shows vanilla data heterogeneity, and the bottom row includes long-tailed federated non-IID settings. (a) FedAvg, (b) FedETF, (c) FedLoGe, and (d) FedYoYo. Our FedYoYo method reaches better neural collapse conditions and achieves the best performance.

$\{(x_i, y_i) \mid 1 \leq i \leq n_k\}$, with n_k representing the number of samples on client k , and x_i, y_i denoting the input data and its corresponding label for each sample i . The model of client k denotes $f_k(x)$. The entire training dataset $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^K$ combines the data from all clients. In vanilla FL, the overall distribution of \mathcal{D} is class-balanced [11, 14, 16], while in our paper, we also consider a more realistic scenario where the global dataset exhibits a long-tailed distribution across the C classes [24]. In this global long-tailed distribution, we assume that classes are ordered by their sample frequencies such that if $i < j$, then $N^i \geq N^j$, where N^i is the total number of samples for class i . For each class c , let n_k^c represent the number of samples of class c on client k , leading to the global sample count of class c as $N^c = \sum_{k=1}^K n_k^c$.

2.2. Augmented Self-bootstrap Distillation

We adopt a self-supervised learning approach similar to BYOL [7], applying different augmentations to local data to capture more robust feature representations. Then we introduce **Augmented Self-bootstrap Distillation**, using the weakly augmented view as the teacher to provide guidance for the strongly augmented view, with the goal of enhanc-

ing feature representation quality. Unlike traditional self-distillation [37], which typically refers to knowledge transfer within a model such as from deeper to shallower layers, our approach distills knowledge across different views for the same model. Also, unlike BYOL, we learn logits instead of features and efficiently use one local model as the teacher itself instead of two online and target models in BYOL. Using logits as the representation can align the representation across clients since the logit space is the unified prediction space; also, we can calibrate the logits by considering both local and global distributions. Specifically, we employ Kullback-Leibler (KL) divergence to align the posterior distributions of the strong and weak augmentations. The distillation loss is then defined by minimizing the KL divergence as follows:

$$\mathcal{L}_{ASD} = \frac{1}{n_k} \sum_{i=1}^{n_k} KL(p(\bar{x}_i) \parallel p(\tilde{x}_i)), \quad (1)$$

where \bar{x}_i and \tilde{x}_i represent the weak and strong augmentations of the same sample. Specifically, we leverage the local model to learn from two augmented instances generated through weak and strong augmentations. For weak augmentations, we apply techniques including RandomCrop, Ran-

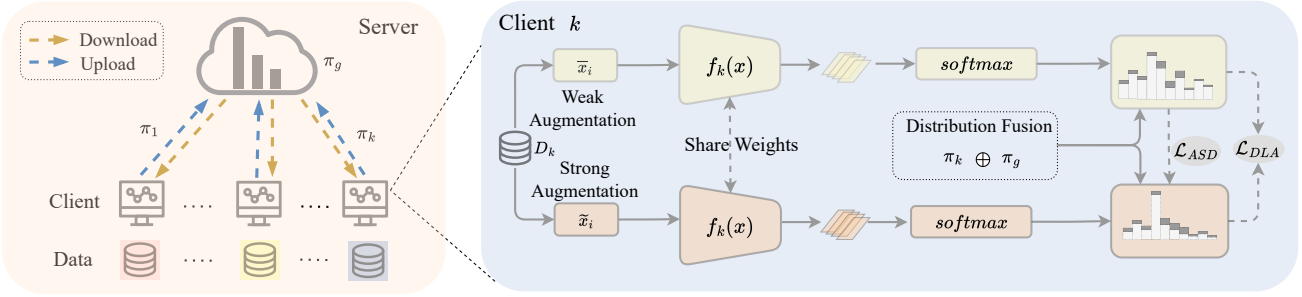


Figure 3. **Overview of our proposed FedYoYo framework.** On the server side, the estimated client distributions are aggregated to obtain an approximate global distribution. On the client side, client k combines its local distribution π_k with the global distribution π_g using EMA updates for balance softmax. During local training, the model $f_k(x)$ processes weak and strong augmentations, using augmented self-bootstrap distillation loss (\mathcal{L}_{ASD}) and distribution-aware logit adjustment loss (\mathcal{L}_{DLA}) to enhance representation learning.

domHorizontalFlip, and RandomRotation. For strong augmentations, we follow [4]. Notably, to avoid introducing noise, only correctly classified samples from the weak augmentation are used as the teacher to guide the learning of the corresponding student, rather than including all samples. Importantly, we adopt Eq. (7) as the adjusted softmax $p(x)$ in Eq. (1) to improve the performance of minority classes.

The ASD technique not only captures richer features by leveraging a bootstrap style but also alleviates client drift by focusing on correctly classified samples. Thus, the local models and local data can be the best teachers for themselves in heterogeneous FL. This targeted guidance helps the local models converge more closely to optimal representations, bridging the performance gap between global and local models. Furthermore, by integrating an adjusted softmax, we effectively mitigate the biases caused by the long-tailed data, resulting in more stable and robust model performance across diverse clients.

2.3. Distribution-aware Logit Adjustment

As mentioned, data heterogeneity often leads to biases in both feature representation and the classifier itself. Such biased classifiers may undermine the effectiveness of representation learning, thereby negatively affecting the local self-bootstrap distillation process. Therefore, it is essential to calibrate the classifier on each client during training to guarantee consistent outputs across all clients. To achieve this, we adopt a balanced softmax for the model output, and the resulting probability distribution is:

$$p(x) = \frac{n^y \exp(f(x, y)/T)}{\sum_{y'=1}^C n^{y'} \exp(f(x, y')/T)}, \quad (2)$$

where n^y is the sample count of class y and T is the temperature coefficient (set to 1.5 in experiments) that controls the smoothness of the output.

In federated long-tailed classification, global data imbalance exacerbates client-side heterogeneity, making sample count an unreliable measure of class representation. Minority classes can occupy large feature space regions despite few samples, while majority classes may show limited

spread due to overlapping features. This reliance on sample count leads to biased class representation and degrades global model performance.

To address this, we propose a global-distribution-aware weight allocation strategy using the Pearson Correlation Coefficient to analyze local sample relationships for more adaptive class weighting. Inspired by previous work [3], this method derives a local weight distribution that better reflects data variability. For class c , we first compute the correlation matrix R_{cb} within each batch b , which measures the pairwise relationships among all samples in the batch. The matrix R_{cb} is defined as:

$$R_{cb} = \left[R_{cb}^{i,j} \right]_{i,j \in b_k^c}, \quad (3)$$

$$R_{cb}^{i,j} = \frac{(h_i - \mu_c)(h_j - \mu_c)^T}{\|h_i - \mu_c\|_2 \cdot \|h_j - \mu_c\|_2}, \quad (4)$$

where b_k^c denotes the set of samples from class c in a single batch, h_i and h_j represent the feature vectors of samples i and j , μ_c is the mean feature vector of class c in the local data, i.e., the class prototype. Next, we compute the effective prior distribution iteratively across batches during training. The effective prior distribution π_k^c is computed as:

$$\pi_k^c = \left\{ \sum_{b=1}^B \frac{1}{a_{cb} R_{cb} a_{cb}^T} \right\}, \quad (5)$$

where B denotes the total number of training batches, and $a_{cb} = \left(\frac{1}{|b_k^c|}, \frac{1}{|b_k^c|}, \dots, \frac{1}{|b_k^c|} \right) \in \mathbb{R}^{1 \times |b_k^c|}$. Finally, we obtain the local prior distribution as $\pi_k = \{\pi_k^1, \pi_k^2, \dots, \pi_k^c\}$.

Additionally, for data heterogeneity, we directly perform logit adjustment using the locally estimated distribution π_k . However, in the case of federated long-tailed learning, the influence of the global distribution must also be considered. To address this, we approximate the global distribution π_g by aggregating client-side π_k at the server by FedAvg, enabling more balanced model training. The final fused distribution π_{mix} is defined as:

$$\pi_{mix} \leftarrow (1 - \gamma) \cdot \pi_g + \gamma \cdot \pi_k, \quad (6)$$

where γ controls the degree of integration, which is discussed in Sec. 3.4.

Then the distribution-fused balanced softmax is expressed as:

$$p(x) = \frac{\pi_{mix}^y \exp(f(x, y)/T)}{\sum_{y'=1}^C \pi_{mix}^{y'} \exp(f(x, y')/T)}. \quad (7)$$

Thus, we propose a distribution-aware logit adjustment loss function, denoted as DLA. The loss function is formulated as follows:

$$\mathcal{L}_{DLA} = -\frac{1}{2n_k} \sum_{i=1}^{2n_k} \log(p(\hat{x}_i)), \quad (8)$$

where \hat{x}_i represents the i -th augmented sample (both weakly and strongly augmented), with $2n_k$ accounting for the two types of augmented samples. Since self-bootstrap distillation involves two augmented views, we apply \mathcal{L}_{DLA} to both views to ensure balanced loss contributions.

2.4. Training

Our method’s local loss function combines two components: distillation loss \mathcal{L}_{ASD} and classifier balancing loss \mathcal{L}_{DLA} . The \mathcal{L}_{ASD} loss, as defined earlier, improves local representation learning. To mitigate classifier bias and its negative impact on feature representation learning, we use the \mathcal{L}_{DLA} loss to adjust the local classifier. During client-side training, \mathcal{L}_{DLA} is applied to learn from hard labels (i.e., the one-hot label), with both augmented batches processed simultaneously by the model. Thus, the total loss function is defined as:

$$\mathcal{L}_{all} = \mathcal{L}_{DLA} + \lambda \mathcal{L}_{ASD}, \quad (9)$$

where λ is a hyperparameter that controls the weight of the distillation loss. The impact of this hyperparameter is discussed in Sec. 3.4. For the server aggregation, we conduct FedAvg to local models.

2.5. Privacy Discussion

Since local distribution estimation occurs entirely client-side, no privacy risk arises in the non-IID scenario. However, privacy concerns may emerge when incorporating global distribution information in federated long-tailed learning—a common challenge in federated learning, not specific to our method, as previous approaches such as FedGrab [32] and FedLoGe [31] have also utilized local data distributions. If privacy protection is required, differential privacy (DP) [6] can be applied by adding noise to uploaded distributions. A comprehensive discussion on federated learning privacy is beyond this work’s scope; thus, we briefly address it here.

3. Experiments

3.1. Experimental Setup

Datasets and models. We first evaluate our model on CIFAR-10/100, where the heterogeneity of the client data is controlled using the concentration parameter α of the Dirichlet distribution (the smaller α , the more heterogeneous data). To further verify the robustness of our method under the more heterogeneous scenarios induced by real-world long-tailed distributions, we conduct experiments on several standard long-tailed datasets: CIFAR-10/100-LT, SVHN-LT, and ImageNet-LT. The imbalance factor (IF) is used to control the degree of imbalance. ImageNet-LT is a long-tailed version of ImageNet, with the largest and smallest categories containing 1,280 and 5 images. For CIFAR-10/100-LT and SVHN-LT, we utilize the ResNet-8 model, while for ImageNet-LT, we employ the ResNet-50 model. The detailed data distribution of CIFAR-10/100-LT is provided in Appendix.A. **Implementation of baseline methods.** We select three categories of state-of-the-art (SOTA) baseline methods for comparison: (1) Heterogeneity-oriented methods (FedProx [14], FedETF [16], FedLC [35], and CCVR [19]) and federated long-tailed methods (CReFF [24], Fed-Grab [32], BalanceFL [27], FedIC [23], RUCR [8], and FedLoGe [31]); (2) Federated distillation methods, including FedDF [18], FedFTG [38], FedGen [42], and DaFKD [29]; (3) Long-tailed methods like τ -norm [10], AREA [3], and LWS [10]. We also present the performance of centralized learning (CL) methods under both heterogeneous and long-tailed settings as an oracle upper bound.

Federated environment and local training. We follow the experimental setup in previous federated long-tailed literature [24]. We train for 300 rounds to reach sufficient convergence. All models are implemented in PyTorch and trained on NVIDIA GeForce 3090 GPUs.

3.2. Comparison with State-of-the-art Methods

Results on vanilla non-IID settings. All results are reported in Tab. 1. Our FedYoYo achieves the best performance under different α . It is notable that our method surpasses FedETF and FedLC in extreme non-IID settings. FedETF is the state-of-the-art method using neural-collapse-inspired classifiers. The results show our method has better generalization than FedETF and Fig. 2 shows that our method can realize better neural collapse optimality than FedETF. FedLC incorporates logit adjustment in FL, but our co-design of self-bootstrap and logit adjustment can reach better performances than logit adjustment solely (i.e., FedLC). Furthermore, our approach successfully reduces the performance gap with the centralized baseline, demonstrating its effectiveness in mitigating the challenges posed by non-IID data distribution.

Table 1. **Top-1 test accuracy (%) of FedYoYo and FL methods with different α .** We compare our method on the vanilla non-IID setting without long-tailed distribution on CIFAR-10/100.

Method	CIFAR-10		CIFAR-100	
	$\alpha=0.01$	$\alpha=0.1$	$\alpha=0.01$	$\alpha=0.1$
Centralized	90.03		68.23	
FedAvg [20]	60.76	72.46	52.46	58.18
FedProx [14]	55.22	70.71	52.91	60.37
CCVR [19]	63.25	74.18	56.95	61.82
FedDF [18]	60.83	79.48	44.49	51.69
FedGen [42]	52.00	71.82	37.72	43.44
FedFTG [38]	51.54	79.07	47.87	53.93
DaFKD [29]	59.28	81.67	47.15	52.17
FedLC [35]	68.49	82.03	58.11	60.44
FedETF [16]	72.31	83.10	58.18	61.87
FedYoYo	81.70	88.82	64.49	67.59

Results on global long-tailed and non-IID settings.

CIFAR-10/100-LT and SVHN-LT: As summarized in Tab. 2, our method achieves the highest test accuracy across all datasets with varying imbalance factors (IFs). Heterogeneity-oriented methods generally perform similarly to FedAvg, as they focus on data heterogeneity but overlook global class imbalance. Imbalance-oriented methods like AREA perform better than FedAvg in certain cases but still lag behind our approach, likely because they primarily address imbalance without accounting for inter-client data heterogeneity. Notably, our method consistently shows significant performance gains over centralized learning with LA loss. **ImageNet-LT:** We further validate our method on the more challenging ImageNet-LT dataset. In Tab. 3, we report accuracies for three class groups: many-shot (over 100 samples per class), medium-shot (20-100 samples), and few-shot (fewer than 20 samples). Our method consistently outperforms others in all groups, particularly in few-shot classes, where it achieves 15.44% accuracy—an 8.04% improvement over the baseline. This showcases the effectiveness of our approach in enhancing few-shot class performance while maintaining strong accuracy in many-shot classes.

3.3. Analysis of FedYoYo

Global-to-local model gap reduction. We visualize the average gains from local training and global aggregation in Appendix.B. Our method significantly reduces the gap between global and local models compared to FedAvg, enhancing the gains from global aggregation. This demonstrates that our approach benefits the global model, enabling faster adaptation to local data, even under the challenging non-IID conditions caused by long-tailed distributions. As local models converge toward optimal consistency, the aggregated model also achieves superior performance. Furthermore, as shown in Fig. 4, our method surpasses FedAvg in creating more balanced across-client

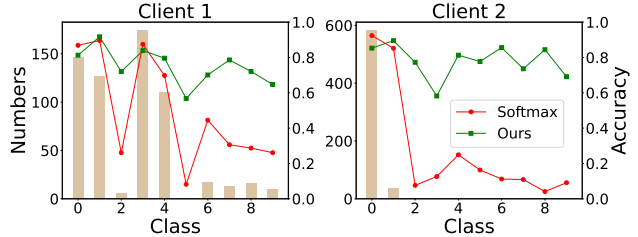


Figure 4. **Comparison of per-class local accuracy after receiving the global model and performing local updates.** Bars represent the local data distribution, while lines indicate accuracy. “Softmax” means FedAvg with vanilla softmax.

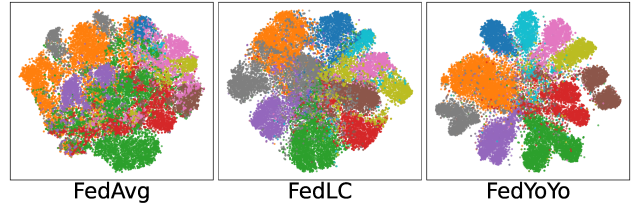


Figure 5. **The t-SNE visualization of feature spaces of FedAvg [20], FedLC [35], and our FedYoYo.**

performance. Overall, our approach exhibits strong adaptability in heterogeneous and long-tailed scenarios, resulting in notable improvements for both global and local models.

Effectiveness of estimated global distribution. Our estimated class distribution closely aligns with the oracle distribution. To further verify this, we tracked the ℓ_2 distance between the estimated and original data distributions across training epochs in Appendix.B. **Feature level analysis.** In Fig. 5, we present a visualization of the feature representations extracted by the global model. Compared to other methods, our approach achieves more compact intra-class features and better inter-class separability. Additionally, we compare the similarity between features of the global model and those of the local models in Appendix.B. Results show that our method effectively reduces the feature discrepancy between local and global models, demonstrating its ability to alleviate client drift and improve local representation learning. **Computation Cost.** We evaluated three methods using ResNet-18 on CIFAR-100-LT (IF=100) in terms of GFLOPs per round. Compared to FedAvg’s 1003.72G and FedGrab’s 2371.36G, our method, FedYoYo, achieves better performance with a more efficient computational cost of 1721.79G.

3.4. Ablation Study

The necessity of distribution fusion. We compare the performance of the real distribution and the estimated distribution and further evaluate the effect of applying a fusion strategy to each. As illustrated in Fig. 6, after applying the fusion strategy, significant improvements are observed in both cases. Notably, under high heterogeneity settings (e.g.,

Table 2. **Top-1 test accuracy (%) of FedYoYo and SOTA methods on CIFAR-10/100-LT, SVHN-LT with different IFs and $\alpha = 0.5$.** The best performance is in bold, and the second is underlined. Red text indicates improved performance compared to the centralized learning with LA loss. \uparrow indicates improved accuracy compared with the underlined (best baseline in each setting).

Method	CIFAR-10-LT			CIFAR-100-LT			SVHN-LT		
	IF=100	IF=50	IF=10	IF=100	IF=50	IF=10	IF=100	IF=50	IF=10
Centralized Methods									
Centralized	68.17	76.15	80.32	38.61	40.30	50.34	84.86	86.55	90.14
Centralized w/ LA Loss [21]	77.27	80.23	86.15	41.57	44.02	56.06	88.17	90.23	93.45
Heterogeneity-oriented FL methods									
FedAvg [20]	58.22	63.47	77.95	31.64	36.14	46.84	81.31	84.40	87.31
FedProx [14]	55.68	60.72	76.64	32.73	35.82	45.77	82.04	86.33	90.13
CCVR [19]	68.13	72.98	81.56	34.73	37.68	48.92	82.34	86.48	91.26
FedLC [35]	69.02	71.93	79.09	31.75	38.36	47.54	81.17	85.23	87.43
FedETF [16]	68.25	72.35	81.60	33.19	37.86	48.71	83.02	87.07	90.86
Imbalance-oriented FL methods									
τ -norm [10]	40.70	41.31	51.66	19.29	20.82	32.51	70.65	74.63	78.58
LWS [10]	37.46	39.82	49.30	18.18	20.10	33.81	71.18	73.25	78.15
AREA [3]	64.33	65.16	78.97	36.34	37.83	48.59	82.87	85.24	90.35
Federated long-tailed methods									
BalanceFL [27]	49.34	54.17	72.03	26.03	29.28	40.16	75.41	79.13	85.16
CReFF [24]	70.51	73.60	79.85	32.90	34.66	43.42	85.47	87.64	91.16
FedIC [23]	66.49	67.55	71.21	33.67	36.74	41.93	84.94	86.82	90.53
Fed-Grab [32]	<u>70.63</u>	<u>75.44</u>	<u>85.21</u>	33.53	44.01	55.87	<u>90.39</u>	<u>91.11</u>	<u>94.56</u>
RUCR [8]	55.32	60.24	75.56	27.61	33.81	41.29	73.45	82.20	87.23
FedLoGe [31]	70.54	74.85	84.54	<u>42.63</u>	<u>47.66</u>	<u>58.36</u>	85.05	87.06	89.68
FedYoYo	81.45 (+4.18)	83.85 (+3.62)	87.94 (+1.79)	46.13 (+4.56)	50.83 (+6.81)	60.16 (+4.10)	91.73 (+3.56)	92.38 (+2.15)	95.10 (+1.65)
	\uparrow 10.82	\uparrow 8.41	\uparrow 2.73	\uparrow 3.50	\uparrow 3.17	\uparrow 1.80	\uparrow 1.34	\uparrow 1.27	\uparrow 0.54

Table 3. **Top-1 test accuracy (%) of FedYoYo and SOTA methods on ImageNet-LT with $\alpha = 0.1$.**

Method	ImageNet-LT			
	Many	Medium	Few	All
FedAvg [20]	34.92	19.18	7.41	23.85
FedProx [14]	34.25	17.06	6.73	22.57
CCVR [19]	36.72	20.24	9.26	25.49
FedLC [35]	36.03	21.14	6.57	23.23
FedETF [16]	35.94	19.91	7.07	23.97
τ -norm [10]	30.81	14.57	5.22	19.58
LWS [10]	37.23	23.4	7.50	25.37
AREA [3]	39.83	23.51	8.53	26.03
BalanceFL [27]	30.63	19.26	7.20	21.87
CReFF [24]	37.61	21.48	10.02	26.91
FedIC [23]	36.22	20.5	9.76	25.71
Fed-Grab [32]	41.16	24.42	14.29	30.56
RUCR [8]	30.47	15.77	5.22	20.06
FedLoGe [31]	40.77	24.16	13.27	29.73
FedYoYo	42.05	25.78	15.44	31.41

$\alpha=0.1$), the estimated distribution with the fusion strategy even outperforms the real distribution, demonstrating that the fusion approach not only mitigates the impact of data heterogeneity but also enhances model generalization. This confirms the value of fusing global and local distributions in strengthening model robustness.

Ablation studies on all components of FedYoYo. In Tab. 4, we present a comprehensive ablation study on the CIFAR-100-LT dataset, evaluating key components: RandAug, ASD, and DLA. RandAug refers to the use of strong

augmentation. A vanilla model without any components reaches an accuracy of 33.34%, similar to FedAvg. When RandAug is combined with ASD, accuracy improves by 8.65%, highlighting ASD’s crucial role in enhancing feature learning and knowledge transfer. Notably, without DLA, many-shot classes see a significant performance boost, but gains for few-shot classes are minimal. When all components are combined, the overall accuracy improves by 12.79%, with gains of 8.42% for medium-shot and 21.13% for few-shot classes. These findings show that DLA mitigates classifier bias and balances the distillation process. Additionally, ASD and DLA reinforce each other, working synergistically to deliver consistent performance improvements across head, medium, and tail classes.

Impact of data augmentation. In this section, we apply various augmentations to the training samples to assess the effectiveness of ASD. As shown in Tab. 5, we compare four augmentation variants in FedYoYo, and the results indicate that the weak-strong self-bootstrap distillation outperforms other augmentations. Moreover, other augmentation methods were examined in Appendix.C.

Influence of fusion coefficient γ . Fig. 7a shows the effect of the fusion coefficient γ on model accuracy across different imbalance factors (IF). A larger γ emphasizes the global distribution. Initially, increasing γ improves accuracy, but excessive values (e.g., 1.0) cause a decline, indicating that relying solely on global or local distributions is suboptimal. Our approach balances global and local information, achieving better overall performance.

Influence of loss weight λ . As shown in Fig. 7b, despite in-

Table 4. Ablation study of our method’s key components on the CIFAR-100-LT dataset with IF = 100 and $\alpha = 0.5$.

RandAug	ASD	DLA	Many	Medium	Few	All
			56.89	31.94	7.50	33.34
✓			60.18	30.49	12.50	35.71
		✓	51.46	34.43	21.13	36.40
✓	✓		65.31	42.71	13.93	41.99
✓		✓	56.03	42.51	27.33	41.69
✓	✓	✓	60.26	47.00	28.63	46.13

Table 5. Top-1 test accuracy (%) accuracy comparison with different augmentation policies on CIFAR100-LT with IF = 100 and $\alpha = 0.5$.

View1(Teacher)	View2(Student)	Accuracy
Weak augmentation	Weak augmentation	33.74
Strong augmentation	Strong augmentation	42.92
Weak augmentation	Strong augmentation	46.13
Strong augmentation	Weak augmentation	41.06

creasing λ from 3.0 to 5.5, the model performance remains relatively stable with minor variations in accuracy across different values. This suggests that the model is not sensitive to the choice of λ , implying that consistent performance can be maintained even with different weights. Therefore, we recommend choosing λ within this range, as it provides stable and reliable performance.

4. Related Works

4.1. Data Heterogeneity in Federated Learning

Our paper focuses on data heterogeneity in federated learning that includes local non-IID heterogeneity and global long-tailed data heterogeneity. Local data heterogeneity causes client drift, which is often mitigated by regularization methods like FedProx [14], SCAF-FOLD [11], and MOON [13]. To tackle classifier bias in non-IID settings, prior works explore strategies such as classifier retraining [19], prototype-based classifier rebalancing [5, 36] and local calibration [35]. However, global long-tailed distributions further exacerbate cross-client heterogeneity, making adaptation more difficult. Recent methods focus on server-side calibration [23, 24] or local model adjustment using global distribution [32], but their effectiveness remains limited, highlighting the need for more robust solutions.

4.2. Feature Representation Learning in FL

Data heterogeneity in federated learning leads to poor feature representations and biased classifiers, causing misalignment between global and client models. Under global long-tailed distributions, classifier bias further worsens this misalignment. Anchor-based methods [33, 41] attempt feature alignment, while neural-collapse-inspired approaches

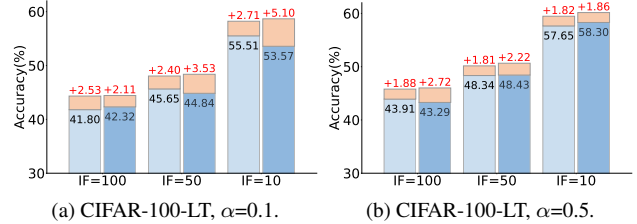


Figure 6. Top-1 test accuracy (%) of our method using different distribution estimation strategies on CIFAR-100-LT with varying α . Light blue bars show results using the ground-truth distribution, while dark blue bars represent the Pearson coefficient-based estimation. Red texts highlight the performance gains from the fused distribution strategy.

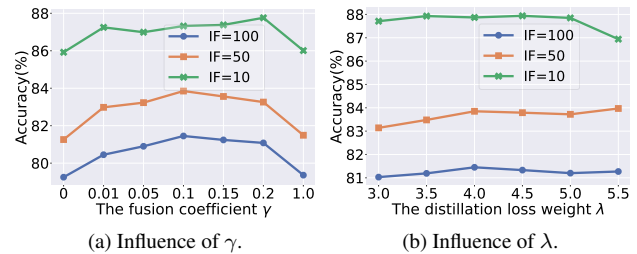


Figure 7. (a) Comparison of different fusion ratios γ . We report the performance across various fusion ratios. (b) Comparison of different values of λ .

employ ETF-based methods [16, 31] or feature regularization [26] to enhance generalization. Others use re-weighting in client aggregation [25] to mitigate bias. However, these methods yield limited improvements in feature representation. Existing self-supervised federated methods typically utilize contrastive learning [2, 43] or bootstrap methods [7], often overlooking minority classes and disproportionately emphasizing majority classes in supervised heterogeneous federated scenarios. In contrast, our method explicitly learns logits as representations and employs logit adjustment guided by distribution information. This strategy effectively enhances minority-class representation, aligns client features, and improves the global model’s generalization and performance.

5. Conclusions

We propose FedYoYo, a novel federated learning method addressing challenges posed by heterogeneous and long-tailed data distributions. FedYoYo integrates Augmented Self-bootstrap Distillation (ASD) and Distribution-aware Logit Adjustment (DLA). ASD employs weakly augmented samples as self-teachers to guide strongly augmented samples, enhancing local feature extraction under client data diversity. DLA leverages both local and global distributions to calibrate logits, providing effective guidance signals for representation learning. Extensive experiments on CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT demonstrate FedYoYo’s state-of-the-art performance, surpassing even centralized baselines in global long-tailed scenarios.

6. Acknowledgements

This study was supported in part by the National Natural Science Foundation of China under Grants 62376233, 62431004, U21A20514, 62372388, and 62466036; in part by the Natural Science Foundation of Fujian Province under Grant 2024J09001; in part by the High-level and Urgently Needed Overseas Talent Programs of Jiangxi Province under Grant 20232BCJ25024; in part by the Zhejiang Provincial Key Research and Development Project under Grant 2023C01043 and Engineering Research Center of Integration and Application of Digital Learning Technology, Ministry of Education; and in part by Xiaomi Young Talents Program.

References

- [1] Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific Reports*, 12(1):1953, 2022. 1
- [2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 8
- [3] Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. Area: Adaptive reweighting via effective area for long-tailed classification. In *The IEEE/CVF International Conference on Computer Vision*, pages 19220–19230, 2023. 4, 5, 7
- [4] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 4
- [5] Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7314–7322, 2023. 8
- [6] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. 5
- [7] Jean-Bastien Grill, Florian Strub, Florent Althé, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 2, 3, 8
- [8] Wenke Huang, Yuxia Liu, Mang Ye, Jun Chen, and Bo Du. Federated learning with long-tailed data via representation unification and classifier rectification. *IEEE Transactions on Information Forensics and Security*, 2024. 5, 7
- [9] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *The International Conference on Learning Representations*, 2022. 2
- [10] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *The International Conference on Learning Representations*, 2020. 5, 7
- [11] 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. 1, 2, 3, 8
- [12] Alexander C Li, Alexei A Efros, and Deepak Pathak. Understanding collapse in non-contrastive siamese representation learning. In *European Conference on Computer Vision*, pages 490–505. Springer, 2022. 2
- [13] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021. 8
- [14] 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. 1, 2, 3, 5, 6, 7, 8
- [15] Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. In *International Conference on Machine Learning*, pages 19767–19788. PMLR, 2023. 1
- [16] Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5319–5329, 2023. 1, 2, 3, 5, 6, 7, 8
- [17] Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Fengyuan Yu, Huabin Zhu, Binhui Yao, Tao Wang, Xiaolin Zheng, and Yanchao Tan. Rethinking the representation in federated unsupervised learning with non-iid data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22841–22850, 2024. 2
- [18] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020. 5, 6
- [19] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021. 2, 5, 6, 7, 8
- [20] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 6, 7
- [21] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *The International Conference on Learning Representations*, 2021. 1, 7

- [22] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 2
- [23] Xinyi Shang, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Fedic: Federated learning on non-iid and long-tailed data via calibrated distillation. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2022. 2, 5, 7, 8
- [24] 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*, pages 2218–2224, 2022. 2, 3, 5, 7, 8
- [25] Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *The International Conference on Learning Representations*, 2021. 8
- [26] Yujun Shi, Jian Liang, Wenqing Zhang, Chuhui Xue, Vincent YF Tan, and Song Bai. Understanding and mitigating dimensional collapse in federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2936–2949, 2023. 8
- [27] Xian Shuai, Yulin Shen, Siyang Jiang, Zhihe Zhao, Zhenyu Yan, and Guoliang Xing. Balancefl: Addressing class imbalance in long-tail federated learning. In *The ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 271–284. IEEE, 2022. 5, 7
- [28] Ha Min Son, Moon-Hyun Kim, Tai-Myoung Chung, Chao Huang, and Xin Liu. Feduv: uniformity and variance for heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5863–5872, 2024. 2
- [29] Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. Dafkd: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20412–20421, 2023. 5, 6
- [30] Chenrui Wu, Haishuai Wang, Xiang Zhang, Zhen Fang, and Jiajun Bu. Spatio-temporal heterogeneous federated learning for time series classification with multi-view orthogonal training. In *Proceedings of the ACM International Conference on Multimedia*, pages 2613–2622, 2024. 1
- [31] Zikai Xiao, Zihan Chen, Liyinglan Liu, Yang Feng, Jian Wu, Wanlu Liu, Joey Tianyi Zhou, Howard Hao Yang, and Zuozhu Liu. Fedloge: Joint local and generic federated learning under long-tailed data. In *The International Conference on Learning Representations*, 2024. 2, 5, 7, 8
- [32] Zikai Xiao, Zihan Chen, Songshang Liu, Hualiang Wang, Yang Feng, Jin Hao, Joey Tianyi Zhou, Jian Wu, Howard Yang, and Zuozhu Liu. Fed-grab: Federated long-tailed learning with self-adjusting gradient balancer. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 7, 8
- [33] Rui Ye, Zhenyang Ni, Chenxin Xu, Jianyu Wang, Siheng Chen, and Yonina C Eldar. Fedfm: Anchor-based feature matching for data heterogeneity in federated learning. *IEEE Transactions on Signal Processing*, 71:4224–4239, 2023. 8
- [34] Xuefei Yin, Yanming Zhu, and Jiankun Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys*, 54(6):1–36, 2021. 1
- [35] 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. 1, 5, 6, 7, 8
- [36] Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 16768–16776, 2024. 8
- [37] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019. 3
- [38] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10174–10183, 2022. 5, 6
- [39] Tuo Zhang, Lei Gao, Chaoyang He, Mi Zhang, Bhaskar Krishnamachari, and A Salman Avestimehr. Federated learning for the internet of things: Applications, challenges, and opportunities. *IEEE Internet of Things Magazine*, 5(1):24–29, 2022. 1
- [40] 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*, 45(9):10795–10816, 2023. 2
- [41] Tailin Zhou, Jun Zhang, and Danny HK Tsang. Fedfa: Federated learning with feature anchors to align features and classifiers for heterogeneous data. *IEEE Transactions on Mobile Computing*, 2023. 8
- [42] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pages 12878–12889. PMLR, 2021. 5, 6
- [43] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. In *The International Conference on Learning Representations*, 2022. 8