Mind the Gap: Confidence Discrepancy Can Guide Federated Semi-Supervised Learning Across Pseudo-Mismatch

Yijie Liu[1,2] Xinyi Shang[3] Yiqun Zhang[4] Yang Lu[1,2*] Chen Gong[5]
Jing-Hao Xue[3] Hanzi Wang[1,2]

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

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

[3]Department of Statistical Science, University College London, United Kingdom

[4]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China

[5]Department of Automation, Shanghai Jiao Tong University, China

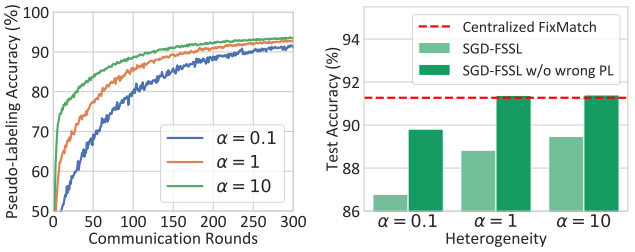
yijieliu@stu.xmu.edu.cn, xinyi.shang.23@ucl.ac.uk, yqzhang@gdut.edu.cn, luyang@xmu.edu.cn,
chen.gong@sjtu.edu.cn, jinghao.xue@ucl.ac.uk, hanzi.wang@xmu.edu.cn

Abstract

*Federated Semi-Supervised Learning (FSSL) aims to leverage unlabeled data across clients with limited labeled data to train a global model with strong generalization ability. Most FSSL methods rely on consistency regularization with pseudo-labels, converting predictions from local or global models into hard pseudo-labels as supervisory signals. However, we discover that the quality of pseudo-label is largely deteriorated by data heterogeneity, an intrinsic facet of federated learning. In this paper, we study the problem of FSSL in-depth and show that (1) heterogeneity exacerbates pseudo-label mismatches, further degrading model performance and convergence, and (2) local and global models' predictive tendencies diverge as heterogeneity increases. Motivated by these findings, we propose a simple and effective method called **S**emi-supervised **A**ggregation for **G**lobally-Enhanced **E**nsemble (SAGE), that can flexibly correct pseudo-labels based on confidence discrepancies. This strategy effectively mitigates performance degradation caused by incorrect pseudo-labels and enhances consensus between local and global models. Experimental results demonstrate that SAGE outperforms existing FSSL methods in both performance and convergence. Our code is available at https://github.com/Jay-Codeman/SAGE.*

1. Introduction

The rapid proliferation of mobile devices and the Internet of Things (IoT) has led to unprecedented growth in distributed data [10, 25]. This shift has created a pressing need for approaches that can leverage decentralized data while preserving user privacy. Federated Learning (FL) addresses this need by enabling collaborative model training directly on edge devices, sharing only model updates rather than raw data [16, 29]. Clients participating in FL typically possess some labeled data and conduct supervised training locally. However, when labeling costs are constrained, only a very small portion of their data may be labeled [13]. To handle this situation, Federated Semi-Supervised Learning (FSSL) has emerged [12, 24], allowing clients to perform Semi-Supervised Learning (SSL) on private data, leveraging a



(a) Pseudo-labeling accuracy. (b) Test accuracy.

Figure 1. **Pseudo-labeling accuracy and test accuracy under varying levels of heterogeneity** (smaller α indicates greater heterogeneity). In each communication round, all clients are trained using FedSGD [29] for one local epoch. From (a), we observe that as heterogeneity increases, pseudo-labeling accuracy declines. In (b), the performance gap between SGD-FSSL and Centralized FixMatch indicates the degradation caused by heterogeneity. We observe that when incorrect pseudo-labels are removed, SGD-FSSL can reach the level of centralized performance. **In short, (a) and (b) show that data heterogeneity can negatively impact both model convergence and final test performance.**

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

large amount of unlabeled data to improve the performance of the global model. Current research assumes data heterogeneity both within and across clients, suggesting that data distributions between clients are different (external imbalance), and within each client, labeled and unlabeled data may differ in distribution (internal imbalance) [2, 5, 50]. In this context, biased labels fail to generalize effectively to unseen unlabeled data.

Existing FSSL methods [8, 20, 44, 45] typically employ consistency regularization algorithms based on pseudo-labeling, using high-confidence predictions from local or global models as pseudo-labels for unlabeled data. However, it cannot completely avoid pseudo-label mismatches even in centralized environments due to the bias of self-training [3]. This inspires us to explore the following questions: *(1) Does heterogeneity exacerbate mismatches of hard pseudo-labels? (2) What extent do incorrect hard pseudo-labels affect FSSL model performance?*

To quantify the impact of incorrect pseudo-labels on model performance, we conduct quick experiments under varying levels of data heterogeneity. As shown in Fig. 1(a), with the increase of data heterogeneity (by the value of α), the accuracy of pseudo-labels under SGD-FSSL (FedSGD+FixMatch) significantly declines with a slower convergence rate, exhibiting a clear deteriorating trend. Fortunately, as shown in Fig. 1(b), SGD-FSSL's accuracy improves substantially once incorrect pseudo-labels are manually removed, approaching the level of centralized FixMatch. These observations suggest that hard pseudo-labels act as aggressive supervisory signals, and their negative impact becomes especially embodied under a high level of data heterogeneity. While it is unrealistic to directly eliminate these incorrect pseudo-labels, we could consider moderately correcting them and thus mitigate their harmful effects as much as possible.

To address the problem of FSSL with the above findings, we propose a new FSSL approach called **SAGE** (**S**emi-supervised **A**ggregation for **G**lobally-enhanced **E**nsemble) to handle the scenario where the clients hold partially labeled data, apply flexible pseudo-label corrections based on the confidence perspective of the global model to mitigate the effect of erroneous hard pseudo-label signals. Firstly, we introduce a collaborative pseudo-label generation mechanism. This approach leverages the global model to guide each client, employing global distribution awareness to compensate for the scarcity of pseudo-labels in local minority classes. Secondly, we propose a dynamic, confidence-driven pseudo-label correction mechanism, inspired by an intriguing observation: as heterogeneity increases, the confidence discrepancy between local and global models gradually widens. Accordingly, we adjust the contributions of local and global hard pseudo-labels to the final pseudo-label based on their confidence discrepancies. This mechanism mitigates the impact of potentially incorrect hard pseudo-labels. Experiments show that SAGE can significantly improve the performance and convergence of the FSSL model.

The main contributions of this paper are as follows:

- This paper reveals an intriguing phenomenon: in FSSL, greater data heterogeneity results in a larger confidence discrepancy between the pseudo-labels generated by local and global models. Accordingly, we offer an explanation for the dynamic relationship between data heterogeneity and confidence discrepancies during training.
- We propose an FSSL method, SAGE, that can evaluate and flexibly correct the pseudo-labels generated by local and global models based on their confidence discrepancies under different levels of data heterogeneity, alleviating the negative impact of aggressive hard pseudo-labeling strategies.
- SAGE outperforms existing FSSL methods in performance and convergence across multiple datasets, demonstrating robustness under varying heterogeneous distributions. Additionally, SAGE can serve as a plugin to enhance the performance of existing FSSL methods.

2. Related Work

2.1. Non-IID in Federated Learning

FL is a distributed machine learning approach enabling collaborative training across clients without sharing raw data [14, 16, 29]. Non-IID data presents a major challenge in FL, as differences in data distributions across clients significantly impact the training of federated models [21, 48, 51]. Numerous studies have explored the mechanisms underlying heterogeneous data's effect and proposed solutions like classifier calibration [23, 28] and client selection [4, 38] to alleviate performance degradation. For example, FedDECORR [36] addresses dimensional collapse in FL due to non-IID data by regularizing local models; FedCal [31] reduces global calibration error by applying client-specific calibration factors, while HiCS-FL [4] estimates statistical heterogeneity by analyzing client updates in the output layer of the network, enabling client clustering and selection. However, these methods have yet to analyze non-IID mechanisms in federated scenarios with semi-supervised learning.

2.2. Semi-Supervised Learning

SSL enhances model generalization by combining limited labeled data and abundant unlabeled data, reducing dependence on labeled samples [33, 49]. Current SSL approaches fall primarily into two categories: consistency regularization and pseudo-labeling strategies. Consistency regularization [1, 34, 40] assumes that a model should yield consistent outputs under different perturbations of the same input, using techniques like perturbation augmentation and contrastive loss to constrain the model. Pseudo-labeling

strategies [11, 18, 32], meanwhile, use the model's own predictions as labels for unlabeled samples. Recent methods like FixMatch [37] efficiently integrate consistency regularization and pseudo-labeling through a lightweight self-training mechanism, with several studies refining this approach [19, 39, 41, 43]. However, pseudo-label generation in self-training methods relies heavily on prediction confidence, and in the heterogeneous setting of FL, the quality of self-generated pseudo-labels can vary greatly, making centralized SSL methods challenging to apply directly to FL scenarios.

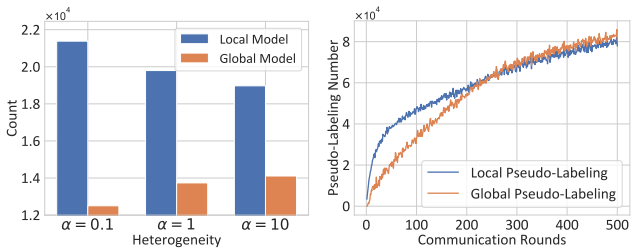
2.3. Federated Semi-Supervised Learning

FSSL settings fall into three categories: (1) Label-at-Server [8, 9, 12, 15, 42], where the server holds some labeled data while clients possess only unlabeled data; (2) Label-at-All-Client [12, 47], where each client contains a small amount of labeled data alongside a large amount of unlabeled data; and (3) Label-at-Partial-Client [20, 24, 26, 27, 46], where only a few clients have fully labeled data, while most have only unlabeled data. Our study focuses on the Label-at-All-Client setting. Recent research [2, 5, 35, 44, 50] builds on FixMatch, focusing on pseudo-label selection or debiasing. However, these methods cannot avoid the impact of incorrect hard pseudo-labels in heterogeneous scenarios.

3. Problem Formulation

This study examines the impact of data heterogeneity on FSSL. We consider both intra-client and inter-client data heterogeneity in FSSL scenarios, with not only external imbalance across clients but also internal imbalance between labeled and unlabeled distributions within each client. Let the set of clients be $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$, where each client C_k trains a local model $f_{l,k}$ parameterized by $\theta_{l,k}$. During each communication round, a subset of online clients $\mathcal{C}_M \subseteq \mathcal{C}$ is randomly selected to participate in training, and the global model f_g aggregates the uploaded model parameters from the clients, obtaining the global parameters θ_g as $\theta_g = \sum_{C_m \in \mathcal{C}_M} w_m \theta_{l,m}$, where w_m represents the weight for client C_m, determined by the proportion of its local dataset size relative to the total number of samples across all participating clients.

Each client C_k maintains a private partially labeled dataset, consisting of labeled data $\mathcal{D}_k^s = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_k^s}$ and unlabeled data $\mathcal{D}_k^u = \{\mathbf{u}_i\}_{i=1}^{N_k^u}$, with $N_k^s \ll N_k^u$, both $\mathcal{D}_k^s$ and $\mathcal{D}_k^u$ demonstrate class imbalance across the label set Y. Specifically, there exists a significant shift between the distribution $Q_k^s(y)$ of the labeled set and the ideal uniform distribution $U = \frac{1}{|Y|}$, i.e., $\mathrm{KL}(Q_k^s(y) \parallel U) \gg 0$. The unlabeled set $\mathcal{D}_k^u$ is assumed to follow the imbalanced distribution $Q_k^u(y)$. For simplicity, we will omit the client index k in the following sections.



(a) Pseudo-label confidence distributions of local and global models.

(b) The number of pseudo-labels under $\alpha = 0.1$.

Figure 2. Differences of the pseudo-labeling ability between local and global models on CIFAR-100. (a) shows the distributions of pseudo-labels with confidence greater than 0.99. As heterogeneity increases (with smaller α), the local and global models exhibit opposite trends. The difference is also reflected in the number of pseudo-labels in (b).

4. Proposed Method

4.1. Preliminary Study

Our goal is to moderately correct potentially incorrect pseudo-labels to mitigate their impact, with the local and global models providing two distinct perspectives on pseudo-labels. To this end, we conduct exploratory experiments to investigate the pseudo-labeling differences between local and global models as data heterogeneity increases. We analyze the confidence distribution of pseudo-labels from both models and track the number of pseudo-labels assigned throughout training. As shown in Fig. 2(a), the confidence of the local model's pseudo-labels shifts more toward the high-confidence region, while the global model exhibits the opposite trend. Additionally, as shown in Fig. 2(b), the local model assigns a higher number of pseudo-labels than the global model in the early stages of training. More exploratory experimental results are shown in Appendix B.1. Based on these experiments, we summarize this phenomenon through two key observations:

Observation 1. *As heterogeneity intensifies, the pseudo-label predictions of the local model grow more confident, while those of the global model become more conservative.*

Observation 2. *The local model exhibits a higher utilization rate of unlabeled data in the early training stages compared to the global model.*

We further analyze the rationale behind these observations to explain why increasing heterogeneity leads to differing predictive tendencies in pseudo-labels between local and global models. The detailed derivation towards the above analysis is provided in Appendix B.2 and B.3. The analysis is as follows:

Remark 1. *The entropy of the local model's predictive distribution, $H(p(y|\mathbf{x}, \mathcal{D}^u))$, is influenced by the entropy of the prior distribution $H(p(y|\mathcal{D}^u))$ and is related to the entropy of the local data distribution $H(Q^u(y))$.*

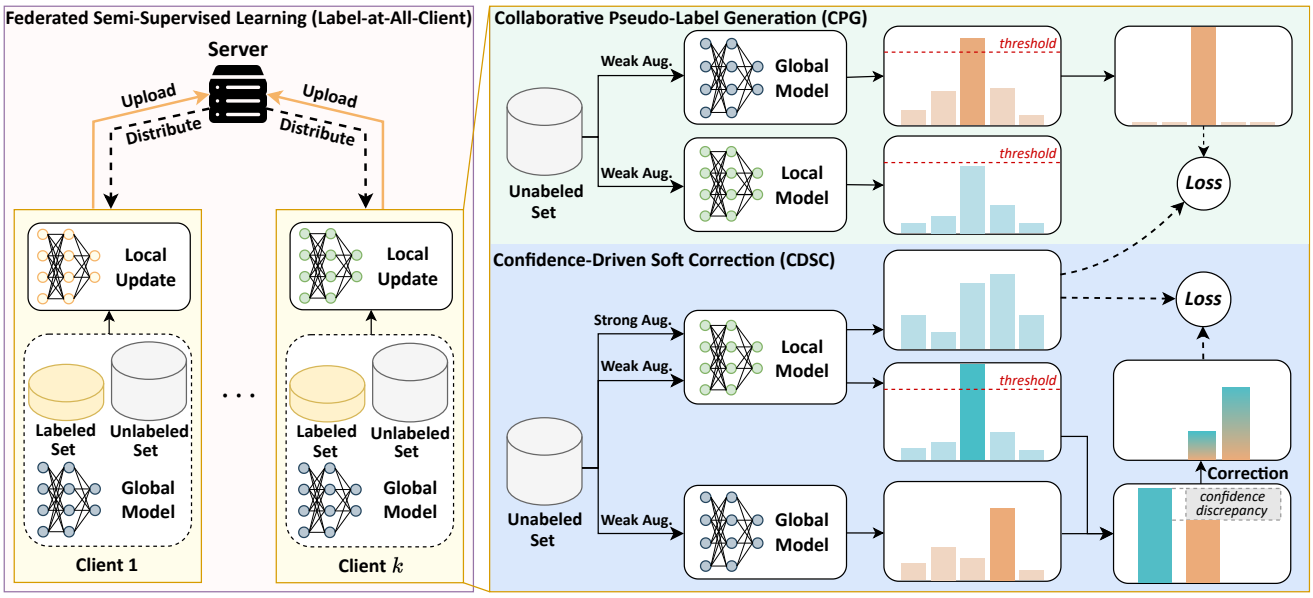


Figure 3. **Framework of the proposed SAGE.** This framework demonstrates the pseudo-labeling strategy of SAGE in the label-at-all-client scenario. The global model's pseudo-labels provide supplementary information when the local model lacks confidence and are dynamically adjusted based on confidence discrepancies between the local and global models.

Remark 2. *The global model's high-confidence predictions increasingly focus on classes with higher consistency across clients, demonstrating more conservative behavior.*

They suggest that the local model tends to overfit when faced with Non-IID data, relying excessively on its imbalanced distribution and being *overly confident* in its predictions, while the global model exhibits a *lack of confidence* as it attempts to create a model that can adapt to the data distribution of all clients.

Based on the above analysis, the pseudo-labeling strategies of the local and global models exhibit substantial discrepancies, offering an opportunity to mitigate the impact of potentially incorrect pseudo-labels by leveraging these discrepancies. To address this, we propose Collaborative Pseudo-Label Generation (CPG) and Confidence-Driven Soft Correction (CDSC), improving unsupervised data utilization while ensuring pseudo-label quality and using flexible pseudo-labels to avoid the radical impacts of hard pseudo-labels. The framework of SAGE approach is shown in Fig. 3.

4.2. Collaborative Pseudo-Label Generation

As discussed above, the pseudo-labeling abilities of the local model f_l and the global model f_g have their respective strengths and weaknesses: f_l is trained on local data, generating a large number of pseudo-labels with high utilization of unsupervised data, but the accuracy of these pseudo-labels cannot be guaranteed. On the other hand, f_g generates fewer pseudo-labels but has a better understanding of the overall data distribution, compensating for the shortcomings of the local model. It can offer robust pseudo-label support to the local model for minority classes, thereby mitigating training errors resulting from the exclusive reliance on local pseudo-labeling strategies. Therefore, we anticipate that integrating the strengths of both models will reduce training errors caused by reliance solely on local pseudo-labels, thereby enhancing the overall pseudo-labeling accuracy.

Therefore, we propose Collaborative Pseudo-Label Generation (CPG) to ensure pseudo-labeling accuracy while enhancing the utilization of unlabeled data. For each unsupervised sample $\mathbf{u} \in \mathcal{D}^u$, we compute the weakly augmented prediction outputs of the local model and the global model, denoted as $p_l(\mathbf{u}) = f_l(\alpha(\mathbf{u}))$ and $p_g(\mathbf{u}) = f_g(\alpha(\mathbf{u}))$, where $\alpha(\mathbf{u})$ represents the weak augmentation (e.g., using only flip-and-shift data augmentation) applied to the unsupervised sample $\mathbf{u}$. We will omit $\mathbf{u}$ in the following text to avoid redundancy. We initially assign pseudo-labels based on the predictions of f_l and f_g:

$$\hat{y} = \begin{cases} \arg\max(p_l) & \text{if } \max(p_l) > \tau, \\ \arg\max(p_g) & \text{else if } \max(p_g) > \tau, \\ \text{N/A} & \text{otherwise,} \end{cases} \tag{1}$$

where τ is the confidence threshold. This strategy, derived from Observation 2, prioritizes obtaining pseudo-labels from the local model and supplements them with predictions from the global model when local confidence is insufficient. This approach ensures pseudo-label quality while further enhancing the utilization of unlabeled data. Building on this, we will further correct pseudo-labels.

4.3. Confidence-Driven Soft Correction

CPG enables local models to maintain a high utilization rate of unlabeled data while compensating for the scarcity of pseudo-labels in local minority classes. Building on this, we further aim to utilize the conservative predictions of the global model to mitigate the impact of incorrect local pseudo-labels. From Observation 1, we can infer that: ***as heterogeneity intensifies, the confidence discrepancy between the local and global models widens.*** This insight suggests that the confidence discrepancy between the local and global model can serve as a measure of local imbalance in the predicted class. Specifically, a larger confidence difference between f_g and f_l indicates a greater discrepancy between the local and global distributions for the locally predicted pseudo-label class of that sample. In such cases, we assign greater weight to the global model to ensure pseudo-label robustness. Conversely, when the confidence discrepancy between f_g and f_l is small, it suggests that the local pseudo-label predictions are reliable. In this scenario, the local model is able to capture the characteristics of the local distribution during the current training iteration. Below, we provide a detailed explanation of the confidence-driven soft correction mechanism.

First, we calculate the confidence difference ΔC between f_l and f_g to characterize the discrepancy between the models:

$$\Delta C = |\max(p_l) - \max(p_g)|. \tag{2}$$

Then, based on ΔC, we dynamically adjust the contribution of each model to the pseudo-labels. We define a dynamic correction coefficient $\lambda(\cdot)$ to regulate the contribution of the local and global model pseudo-labels. As ΔC increases, we should decrease the influence of local pseudo-labels and rely more on the conservative predictions of the global model. Therefore, the correction coefficient takes the form of an exponential decay:

$$\lambda = \exp(-\kappa \cdot \Delta C), \tag{3}$$

where κ is a hyperparameter that controls the sensitivity of the correction coefficient.

Next, based on λ, we perform linear interpolation between the predictions of f_l and f_g. We first convert the local pseudo-label and the global predicted class into one-hot form:

$$\delta_l = \text{one-hot}(\arg\max(p_l)), \tag{4}$$

$$\delta_g = \text{one-hot}(\arg\max(p_g)). \tag{5}$$

Then the corrected local pseudo-label is obtained through linear interpolation of them:

$$\tilde{y} = \lambda \cdot \delta_l + (1 - \lambda) \cdot \delta_g. \tag{6}$$

Based on this linear correction, when the confidence predictions of f_l and f_g are more consistent, we rely more on f_l's prediction; when there is a larger discrepancy, we rely more on f_g's prediction. The final pseudo-label $\hat{y}$ can be expressed as:

$$\hat{y} = \begin{cases} \tilde{y} & \text{if } \max(p_l) > \tau, \\ \arg\max(p_g) & \text{else if } \max(p_g) > \tau, \\ \text{N/A} & \text{otherwise.} \end{cases} \tag{7}$$

Through dynamic and flexible correction, CDSC mitigates the radical impact of hard pseudo-labels.

4.4. Loss Functions

For a batch of unlabeled samples B_u, we use KL divergence to compute the unsupervised loss between the corrected soft pseudo-label and the local model's strongly augmented prediction for the sample $\mathbf{u}$, denoted as $p_l(\mathcal{A}(\mathbf{u}))$:

$$L_u = \frac{1}{|B_u|} \sum_{\mathbf{u} \in B_u} \text{KL}\left(p_l(\mathcal{A}(\mathbf{u})) \,\middle\|\, \hat{y}\right), \tag{8}$$

where $\mathcal{A}(\mathbf{u})$ refers to RandAugment with random magnitude [6]. For a batch of labeled samples B_s, we calculate the cross entropy between the local model's predictions and the ground-truth labels: $L_s = \frac{1}{|B_s|} \sum_{\mathbf{x} \in B_s} \mathcal{L}_{CE}(p_l(y|\mathbf{x}, \mathbf{y}))$, where $\mathcal{L}_{CE}$ is the cross-entropy loss. The final loss is a combination of supervised and unsupervised loss:

$$\mathcal{L} = L_s + \mu_u \cdot L_u. \tag{9}$$

We follow the setup in FixMatch [37] where L_s and L_u have the same weight, i.e., $\mu_u = 1$.

The process of SAGE is presented in Algorithm 1 in Appendix A. Using the CPG and CDSC components, SAGE leverages the high utilization of the local model and the balanced distribution of the global model, enabling a "safer" utilization of unlabeled data. This approach mitigates the harmful effects of erroneous hard pseudo-labels and enhances the consensus between local and global models.

5. Experiments

5.1. Experimental Setup

Datasets. We evaluated the SAGE method on the CIFAR-10, CIFAR-100, SVHN, and CINIC-10 datasets [7, 17, 30]. For each dataset, we divided the labeled and unlabeled datasets per class with label proportions of 10% and 20%. We focus on evaluating the performance of methods under more challenging conditions of heterogeneous data. In line with previous work in the FSSL field [2, 5, 50], we simulated both inter-client and intra-client imbalances by sampling labeled and unlabeled data from a Dirichlet distribution $\text{Dir}(\alpha)$ and allocating them equally to each client. We simulated three levels of heterogeneity: $\alpha \in \{0.1, 0.5, 1\}$, A smaller α

Table 1. Experimental results on CIFAR-10, CIFAR-100, SVHN and CINIC-10 under 10% label. Bold text indicates the best result, while underlined text indicates the second-best result. The last row presents the improvement of SAGE over existing methods.

Methods	CIFAR-10			CIFAR-100			SVHN			CINIC-10		
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$
SL methods												
FedAvg	69.60	68.88	69.39	34.08	33.21	35.31	82.40	83.40	78.60	57.17	60.09	61.54
FedProx	68.58	69.53	68.00	34.20	34.07	34.88	81.67	83.77	83.77	58.05	60.71	62.82
FedAvg-SL	90.46	91.24	91.32	67.98	68.83	69.10	94.11	94.41	94.40	77.82	80.42	81.29
SSL methods												
FixMatch-LPL	82.98	84.36	84.69	49.32	49.67	49.55	89.68	91.33	91.91	68.02	70.67	72.69
FixMatch-GPL	84.56	86.05	86.66	48.96	51.80	52.19	90.50	91.94	92.31	71.67	73.26	74.80
FedProx+FixMatch	84.60	85.49	86.95	48.42	48.51	49.33	90.46	91.36	91.25	68.62	70.67	72.69
FedAvg+FlexMatch	84.21	86.00	86.57	49.91	51.39	51.79	52.58	55.59	60.50	69.20	71.87	73.42
FSSL methods												
FedMatch [12]	75.35	77.86	78.00	32.23	31.49	35.75	88.63	89.20	89.23	51.94	56.27	70.22
FedLabel [5]	62.85	79.46	79.17	50.88	52.21	52.38	89.31	91.51	91.16	67.64	70.56	72.80
FedLoke [44]	83.32	82.22	81.87	39.29	40.46	39.96	89.94	90.00	89.45	59.03	61.60	63.21
FedDure [2]	84.60	85.88	87.34	48.27	51.09	50.79	92.87	93.49	94.19	70.86	73.37	74.89
FedDB [50]	83.99	85.28	87.49	48.43	50.11	51.55	92.56	93.00	93.14	69.44	72.60	73.61
SAGE (ours)	**87.05**	**88.05**	**89.08**	**54.18**	**55.82**	**56.06**	**93.85**	**94.27**	**94.65**	**74.59**	**75.74**	**76.68**
	↑ 2.45	↑ 2.00	↑ 1.59	↑ 3.3	↑ 3.61	↑ 3.68	↑ 0.98	↑ 0.78	↑ 0.46	↑ 2.92	↑ 2.37	↑ 1.79

value indicates higher data heterogeneity. The specific data distribution is shown in the visualization of Fig. 16 in Appendix E. For all methods, we follow the FixMatch setup and add labeled samples without labels into the unlabeled dataset to enhance sample diversity in the unsupervised dataset. We compared the following methods in our experiments:

- **FL methods (FedAvg [29], FedProx [22], FedAvg-SL):** For FedAvg and FedProx, models are trained via supervised federated learning using only the labeled dataset. FedAvg-SL denotes the standard federated training of FedAvg on the fully labeled dataset, indicating the ideal upper bound.
- **Vanilla combinations:** These methods simply combine SSL methods with FL methods. Notably, for FedAvg+FixMatch, we further subdivided it into "local model pseudo-labeling" and "global model pseudo-labeling" to illustrate differences in pseudo-labeling capabilities between local and global models, abbreviated as **FixMatch-LPL** and **FixMatch-GPL**.
- **FSSL methods:** SAGE is compared with state-of-the-art FSSL methods, including FedMatch [12], FedLoke [44], FedLabel [5], FedDure [2], and FedDB [50]. All of them follow the Label-at-All-Client scenario.

Implementation Details. We assume a total of $|\mathcal{C}| = 20$ clients participating in FL, with $|\mathcal{C}_M| = 8$ clients randomly selected each round for global training. ResNet-8 serves as the backbone network locally, with the number of local epochs set to $E = 5$ and the local learning rate set to $\gamma = 0.1$. Except for FlexMatch, the pseudo-label confidence threshold for all other methods is set to $\tau = 0.95$. Unless otherwise specified, SAGE follows the FixMatch setup in this section. All experiments are conducted three times, with standard deviations shown as error bars in the figures.

5.2. Performance Comparison

Tab. 1 presents the accuracy of various methods across different datasets and under Non-IID settings with 10% label. Under inter-client and intra-client imbalances, FixMatch-GPL outperforms FixMatch-LPL because the global model's pseudo-label generation is unaffected by local data distributions. Most existing FSSL methods based on hard pseudo-labels provide limited performance improvements and, in some cases, perform worse than the vanilla FixMatch method on certain datasets. In contrast, SAGE significantly mitigates the impact of potentially incorrect pseudo-labels by integrating local and global model predictions, *achieving the highest test accuracy across all datasets, with more substantial improvements as the heterogeneity increases.* On the SVHN dataset, SAGE even reaches the performance of fully labeled FedAvg-SL. We attribute this improvement to the enhanced generalization brought by data augmentation. Other labeling ratios are provided in Tab. 8 in Appendix D.2, *where SAGE also achieves the best performance.*

5.3. Convergence Rate

As shown in Fig. 4 and Tab. 2, SAGE significantly speeds up the convergence rate and test accuracy on the CIFAR-100 dataset when $\alpha = 1$ (Other heterogeneous scenarios are simi-

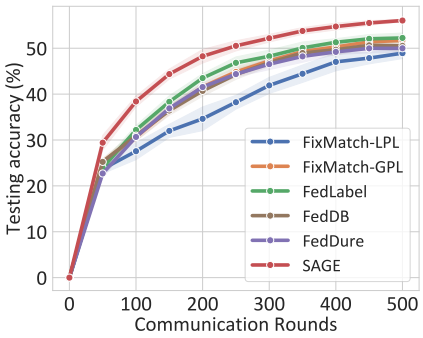


Figure 4. Convergence curves of SAGE and other baselines on CIFAR-100 with $\alpha = 1$.

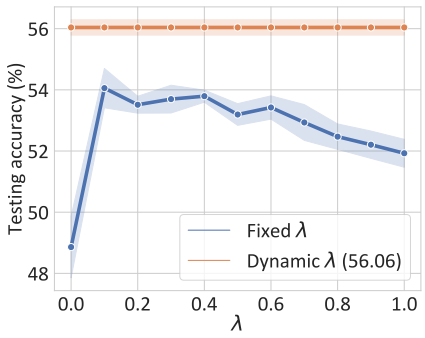


Figure 5. Ablation on Dynamic Correction Coefficient λ.

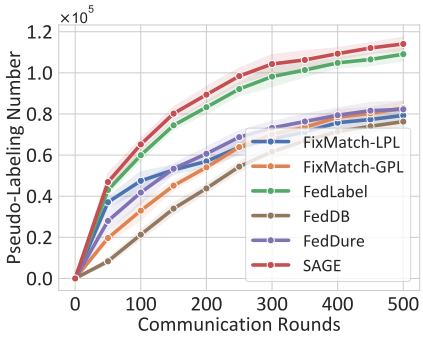


Figure 6. Comparison of pseudo-label counts on CIFAR-100.

Table 2. Comparison of convergence rates between SAGE and other baseline methods with $\alpha = 1$.

Acc.	30%		40%		50%	
Method	Round↓	Speedup↑	Round↓	Speedup↑	Round↓	Speedup↑
LPL	118	×1.00	267	×1.00	527	×1.00
GPL	94	×1.26	183	×1.46	390	×1.35
FedLabel	91	×1.30	164	×1.63	341	×1.55
FedDB	103	×1.15	237	×1.13	418	×1.26
FedDure	95	×1.24	182	×1.47	450	×1.17
SAGE	**56**	**×2.11**	**112**	**×2.38**	**242**	**×2.18**

Table 3. Module ablation studies on CPG and CDSC.

CPG	CDSC	CIFAR100			CINIC10		
		$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$
		49.32	49.67	49.55	68.02	70.67	72.69
✓		52.25	53.85	53.50	72.19	73.14	73.91
	✓	52.43	53.17	53.48	72.83	73.22	74.10
✓	✓	**54.18**	**55.82**	**56.06**	**74.59**	**75.74**	**76.68**

lar and are provided in Appendix D.1). Compared to baseline and existing FSSL methods, *SAGE achieves higher accuracy within fewer communication rounds.* Existing FSSL methods based on hard pseudo-label strategies amplify the impact of incorrect pseudo-labels, leading to greater divergence of local models under non-IID conditions. In contrast, SAGE dynamically corrects pseudo-labels using the global model, establishing stronger consensus between local and global models, thereby accelerating model convergence in the early stages of training.

5.4. SAGE as a Plug-in Approach

The CPL and CDSC components of SAGE function as pseudo-labeling mechanisms agnostic to local semi-supervised training specifics, allowing integration as plugins into hard pseudo-labeling-based SSL and FSSL methods. As shown in Tab. 4, *SAGE improves the performance of existing methods.* This is especially beneficial for FlexMatch, which, due to its strategy of dynamically adjusting class thresholds, is prone to overfitting under class imbalance, a problem exacerbated in non-IID settings. SAGE mitigates this issue by incorporating global information into the pseudo-labeling strategy, resulting in significant performance improvements for FlexMatch on the SVHN dataset.

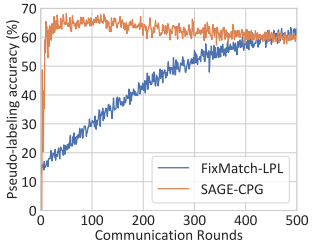
Table 4. Performance gains brought by SAGE as a plugin to other baseline methods.

Methods	CIFAR-100			SVHN		
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$
FixMatch	49.32	49.67	49.55	90.46	91.36	91.25
+SAGE	54.18	55.82	56.06	93.85	94.27	94.65
	↑ 4.86	↑ 6.15	↑ 6.51	↑ 3.39	↑ 2.91	↑ 3.40
FlexMatch	49.91	51.39	51.79	52.58	55.59	60.50
+SAGE	49.84	51.41	52.06	93.36	94.26	93.86
	↓ 0.07	↑ 0.02	↑ 0.27	↑ 40.78	↑ 38.67	↑ 33.36
FedDure	48.27	51.09	50.79	92.87	93.49	94.19
+SAGE	54.13	56.23	55.84	93.96	94.11	94.31
	↑ 5.86	↑ 5.14	↑ 5.05	↑ 1.09	↑ 0.62	↑ 0.12
FedDB	48.43	50.11	51.55	92.56	93.00	94.14
+SAGE	48.33	50.27	51.84	92.51	93.16	93.42
	↓ 0.10	↑ 0.16	↑ 0.29	↓ 0.05	↑ 0.16	↑ 0.28

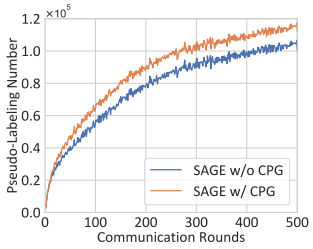
5.5. Ablation Study

In this section, we conduct an in-depth ablation study to demonstrate the contributions of CPG and CDSC within SAGE. More ablation studies on hyperparameter tuning and experiments under different heterogeneity are provided in Appendix. C.

Effectiveness of Components. We first validated the contributions of CPG and CDSC through ablation experiments. FedAvg+FixMatch-LPL, the vanilla combination of FedAvg with FixMatch, served as the baseline method. Experiments were conducted on client data with different levels of data heterogeneity $\alpha = \{0.1, 0.5, 1\}$ to assess component effectiveness. As shown in Tab. 3, *each component consistently enhances model performance under different levels of heterogeneity.* With both CPG and CDSC included, SAGE achieves the best performance gain.

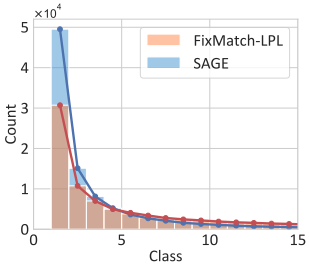


(a) Comparison of pseudo-labeling Acc. between CPG and baseline.

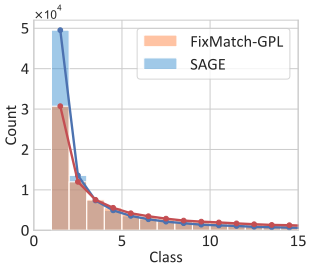


(b) The increase in pseudo-labels generated by CPG in SAGE.

Figure 7. In-depth ablation of CPG on CIFAR-100. CPG significantly increases the utilization of unlabeled data of SAGE while ensuring pseudo-labeling accuracy.



(a) Ranking in global predictions.



(b) Ranking in local predictions.

Figure 8. Consensus ablation between local and global models. (a) displays the ranking statistics of the local model's pseudo-labels within the global model's class predictions, while (b) displays the ranking statistics of the global model's pseudo-labels within the local model's class predictions.

Pseudo-label Gains from CPG. We monitor the number of pseudo-labels generated by SAGE and the baselines throughout training. As shown in Fig. 6, with the enhancement provided by CPG, *SAGE consistently maintains a lead in pseudo-label count*, a key factor in SAGE's performance improvement. We conducted a further analysis of the performance gains from CPG. As shown in Fig. 7, compared to a single local pseudo-labeling strategy, CPG generates high-accuracy pseudo-labels early in training. With the assistance of the global model, *CPG effectively compensates for the scarcity of pseudo-labels in local minority classes, further enhancing the utilization of unlabeled data*.

Dynamic Correction Coefficient λ. In CDSC, the correction coefficient $\lambda(x)$ quantifies prediction discrepancy between local and global models, balancing the confident predictions of the local model with the conservative predictions of the global model. To evaluate its effectiveness, we compared the dynamic coefficient against fixed values of λ (ranging from 0 to 1). When $\lambda = 0$, the method reduces to FixMatch-LPL, relying only on local pseudo-labels; when $\lambda = 1$, it relies solely on global pseudo-labels, as in FixMatch-GPL. Experimental results in Fig. 5 demonstrate that *regardless of the fixed value of λ, the model's performance surpasses both FixMatch-LPL and FixMatch-GPL, but does not achieve the effectiveness of the dynamic λ*. This finding suggests that assigning a greater global weight to samples with larger confidence discrepancies can more effectively mitigate the impact of potentially incorrect pseudo-labels and thus improve model performance. More ablation studies on λ are provided in Appendix C.2.

CDSC Enhances Consensus Between Global and Local. As stated in Remark 1, existing FSSL methods based on hard pseudo-labels cause local models to fit local biased distributions more aggressively, amplifying the discrepancy between global and local models. Fig. 8 presents a histogram of predicted class rankings, demonstrating the improvement in predictive consensus achieved by SAGE. Taking Fig. 8(a) as an example, after applying SAGE the pseudo-label predictions of the local model tend to rank higher within the global model's class predictions. Similarly, Fig. 8(b) exhibits that the predictions of the global model exhibit the same trend. This indicates that *SAGE effectively reduces prediction discrepancies between local and global models, thereby enhancing their consensus and accelerating the model convergence.*

6. Conclusion

In this study, it was initially observed that increasing heterogeneity leads to pseudo-label mismatches in FSSL, which subsequently affect model performance and convergence. Another intriguing phenomenon was discovered: as heterogeneity increases, the confidence discrepancy between the local and global models expands.We analyzed the underlying rationale and, based on this observation, proposed a new approach called SAGE.SAGE leverages confidence discrepancies for flexible pseudo-label correction, enhancing the utilization of unlabeled data, mitigating the adverse effects of incorrect pseudo-labels, and strengthening the consensus between local and global models. In future work, we aim is to extend the applicability of SAGE to ensure robust performance across different FSSL scenarios, including Label-at-Partial-Client and Label-at-Server settings.Client and Label-at-Server settings.

Acknowledgements

This study was supported in part by the National Natural Science Foundation of China under Grants 62376233, 62431004, 62476063, U21A20514, 62336003 and 12371510; in part by the Natural Science Foundation of Fujian Province under Grant 2024J09001; and in part by Xiaomi Young Talents Program.

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[2] Sikai Bai, Shuaicheng Li, Weiming Zhuang, Jie Zhang, Kunlin Yang, Jun Hou, Shuai Yi, Shuai Zhang, and Junyu Gao. Combating data imbalances in federated semi-supervised learning with dual regulators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10989–10997, 2024. 2, 3, 5, 6

[3] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning. *Advances in Neural Information Processing Systems*, 35:32424–32437, 2022. 2

[4] Huancheng Chen and Haris Vikalo. Heterogeneity-guided client sampling: Towards fast and efficient non-iid federated learning. *Advances in Neural Information Processing Systems*, 37:65525–65561, 2024. 2

[5] Yae Jee Cho, Gauri Joshi, and Dimitrios Dimitriadis. Local or global: Selective knowledge assimilation for federated learning with limited labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17087–17096, 2023. 2, 3, 5, 6

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 5

[7] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018. 5

[8] Enmao Diao, Jie Ding, and Vahid Tarokh. Semifl: Semi-supervised federated learning for unlabeled clients with alternate training. *Advances in Neural Information Processing Systems*, 35:17871–17884, 2022. 2, 3

[9] Chaoyang He, Zhengyu Yang, Erum Mushtaq, Sunwoo Lee, Mahdi Soltanolkotabi, and Salman Avestimehr. SSFL: Tackling label deficiency in federated learning via personalized self-supervision. *arXiv preprint arXiv:2110.02470*, 2022. 3

[10] Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019. 1

[11] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021. 3

[12] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. *arXiv preprint arXiv:2006.12097*, 2021. 1, 3, 6

[13] Yilun Jin, Xiguang Wei, Yang Liu, and Qiang Yang. Towards utilizing unlabeled data in federated learning: A survey and prospective. *arXiv preprint arXiv:2002.11545*, 2020. 1

[14] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019. 2

[15] Taehyeon Kim, Eric Lin, Junu Lee, Christian Lau, and Vaikkunth Mugunthan. Navigating data heterogeneity in federated learning: a semi-supervised federated object detection. *Advances in Neural Information Processing Systems*, 36:2074–2096, 2023. 3

[16] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016. 1, 2

[17] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009. 5

[18] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 3

[19] Hyuck Lee and Heeyoung Kim. Cdmad: Class-distribution-mismatch-aware debiasing for class-imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23891–23900, 2024. 3

[20] Ming Li, Qingli Li, and Yan Wang. Class balanced adaptive pseudo labeling for federated semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16292–16301, 2023. 2, 3

[21] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Transactions on Signal Processing*, 37(3): 50–60, 2020. 2

[22] 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

[23] 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. 2

[24] Xiaoxiao Liang, Yiqun Lin, Huazhu Fu, Lei Zhu, and Xiaomeng Li. Rscfed: Random sampling consensus federated semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10154–10163, 2022. 1, 3

[25] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020. 1

[26] Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng. Federated semi-supervised medical image classification via inter-client relation matching. In *Medical Image Computing and Computer Assisted Intervention*, pages 325–335, 2021. 3

[27] Yuzhi Liu, Huisi Wu, and Jing Qin. Fedcd: Federated semi-supervised learning with class awareness balance via dual teachers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3837–3845, 2024. 3

[28] 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

[29] 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, 2017. 1, 2, 6

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems Workshops*, page 4. Granada, 2011. 5

[31] Hongyi Peng, Han Yu, Xiaoli Tang, and Xiaoxiao Li. Fedcal: Achieving local and global calibration in federated learning via aggregated parameterized scaler. *arXiv preprint arXiv:2405.15458*, 2024. 2

[32] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 3

[33] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 28, 2015. 2

[34] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016. 2

[35] Xinyi Shang, Gang Huang, Yang Lu, Jian Lou, Bo Han, Yiuming Cheung, and Hanzi Wang. Federated semi-supervised learning with annotation heterogeneity. *arXiv preprint arXiv:2303.02445*, 2023. 3

[36] Yujun Shi, Jian Liang, Wenqing Zhang, Vincent YF Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. *arXiv preprint arXiv:2210.00226*, 2022. 2

[37] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 3, 5

[38] Minxue Tang, Xuefei Ning, Yitu Wang, Jingwei Sun, Yu Wang, Hai Li, and Yiran Chen. Fedcor: Correlation-based active client selection strategy for heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10102–10111, 2022. 2

[39] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022. 3

[40] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. 2

[41] Lihe Yang, Zhen Zhao, Lei Qi, Yu Qiao, Yinghuan Shi, and Hengshuang Zhao. Shrinking class space for enhanced certainty in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16187–16196, 2023. 3

[42] Mingzhao Yang, Shangchao Su, Bin Li, and Xiangyang Xue. Exploring one-shot semi-supervised federated learning with pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16325–16333, 2024. 3

[43] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3

[44] Chao Zhang, Fangzhao Wu, Jingwei Yi, Derong Xu, Yang Yu, Jindong Wang, Yidong Wang, Tong Xu, Xing Xie, and Enhong Chen. Non-iid always bad? semi-supervised heterogeneous federated learning with local knowledge enhancement. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 3257–3267, 2023. 2, 3, 6

[45] Jie Zhang, Xiaosong Ma, Song Guo, and Wenchao Xu. Towards unbiased training in federated open-world semi-supervised learning. In *Proceedings of the International Conference on Machine Learning*, pages 41498–41509, 2023. 2

[46] Yonggang Zhang, Zhiqin Yang, Xinmei Tian, Nannan Wang, Tongliang Liu, and Bo Han. Robust training of federated models with extremely label deficiency. In *Proceedings of the International Conference on Learning Representations*, 2024. 3

[47] Chen Zhao, Zhipeng Gao, Qian Wang, Zijia Mo, and Xinlei Yu. Fedgan: A federated semi-supervised learning from non-iid data. *International Conference on Wireless Algorithms, Systems, and Applications*, pages 181–192, 2022. 3

[48] 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

[49] Zhi-Hua Zhou and Ming Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541, 2005. 2

[50] Guogang Zhu, Xuefeng Liu, Xinghao Wu, Shaojie Tang, Chao Tang, Jianwei Niu, and Hao Su. Estimating before debiasing: A bayesian approach to detaching prior bias in federated semi-supervised learning. *arXiv preprint arXiv:2405.19789*, 2024. 2, 3, 5, 6

[51] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021. 2