Non-IID always Bad? Semi-Supervised Heterogeneous Federated Learning with Local Knowledge Enhancement

Chao Zhang
University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
zclfe00@gmail.com

Fangzhao Wu
Microsoft Research Asia
wufangzhao@gmail.com

Jingwei Yi
University of Science and Technology of China
yjw1029@mail.ustc.edu.cn

Derong Xu
University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
derongxu@mail.ustc.edu.cn

Yang Yu
University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
yflyl613@mail.ustc.edu.cn

Jindong Wang
Microsoft Research Asia
jindong.wang@microsoft.com

Yidong Wang
Microsoft Research Asia
yidongwang37@gmail.com

Tong Xu*
University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
tongxu@ustc.edu.cn

Xing Xie
Microsoft Research Asia
xingx@microsoft.com

Enhong Chen
University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
cheneh@ustc.edu.cn

ABSTRACT

Federated learning (FL) is important for privacy-preserving services by training models without collecting raw user data. Most FL algorithms assume all data is annotated, which is impractical due to the high cost of labeling data in real applications. To alleviate the reliance on labeled data, semi-supervised federated learning (SSFL) has been proposed to utilize unlabeled data on clients to improve model performance. However, most existing methods either have privacy issues which share models trained on other clients, or generate pseudo-labels for unlabeled local datasets with the global model, which is usually biased towards the global data distribution. The latter may lead to sub-optimal accuracy of pseudo-labels, due to the gap between the local data distribution and the global model, especially in non-IID settings. In this paper, we propose a semi-supervised heterogeneous federated learning method with local knowledge enhancement, called FedLoKe, which aims to train an accurate global model from both labeled and unlabeled local data with non-IID distributions. Specifically, in FedLoKe, the server maintains a global model to capture global data distribution, and each client learns a local model to capture local data distribution. Since the distribution captured by the local model is aligned with the local data distribution, we utilize it to generate high-accuracy pseudo-labels of the unlabeled dataset for global model training. To prevent the local model from severely overfitting local labeled data, we further use the exponential moving average and apply the global model to generate pseudo-labels for local modeling training. Experiments on four datasets show the effectiveness of FedLoKe. Our code is available at: https://github.com/zcfinal/FedLoKe.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

Federated Learning; Semi-Supervised; Heterogeneity; Pseudo-Labeling

ACM Reference Format:
Chao Zhang, Fangzhao Wu, Jingwei Yi, Derong Xu, Yang Yu, Jindong Wang, Yidong Wang, Tong Xu, Xing Xie, and Enhong Chen. 2023. Non-IID always Bad? Semi-Supervised Heterogeneous Federated Learning with Local Knowledge Enhancement. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614991

1 INTRODUCTION

Growing use of mobile and edge devices leads to increasing private user data [43]. Some centralized machine learning applications collect the private data of clients and train models at a central server, which raises privacy concerns [14, 27]. To tackle this problem, federated learning (FL) [38] has been proposed to train models

*Corresponding authors.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614991

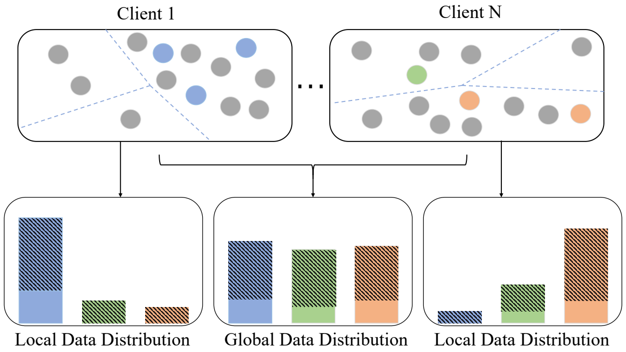


Figure 1: Heterogeneity makes pseudo-labeling easier due to consistent and imbalanced distributions of local labeled and unlabeled dataset. The grey and colorful circles mean unlabeled and labeled data, respectively. Circles with the same color are data from the same class. Blue dotted lines show the decision boundary of three classes. The colorful histograms mean different distributions, and the shaded area means the proportion of unlabeled data.

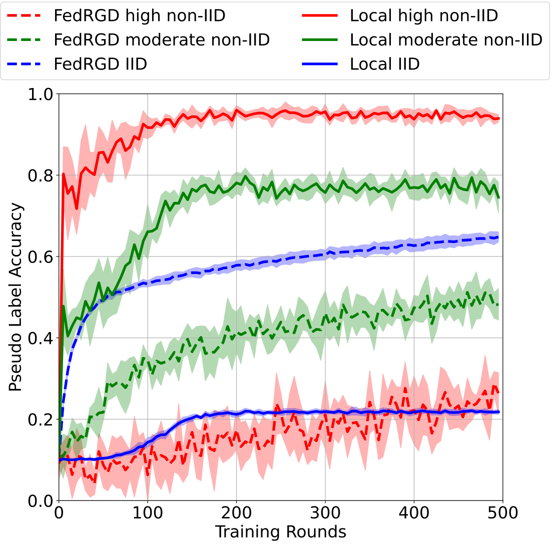


Figure 2: Pseudo-label accuracy of local models only using local labeled data and FedRGD [63], which is a SOTA SSFL method, under different non-IID degrees on CIFAR-10. When the heterogeneity increases, local models achieve better pseudo-label accuracy.

via gradient or parameter aggregation while keeping the privacy-sensitive data from leaving the device of the client, which has been applied in many applications [5, 24, 59]. Most FL methods assume clients possess fully annotated data [28, 30, 45]. However, this assumption is unrealistic. Since massive labeled data collection is expensive, most real-world data on user devices are unlabeled [15].

Semi-supervised federated learning (SSFL) methods have been proposed to improve global model performance using clients' unlabeled data [3, 9, 15, 63]. Along this line, similar to supervised FL, the distributions of clients' data in SSFL might be highly heterogeneous, i.e., non-IID (independently and identically distributed) [15, 20, 32]. To address the non-IID problem, some existing SSFL methods send models trained on some clients to other clients [15, 20, 32] for consistent pseudo-labeling. However, these methods face serious privacy issues [51, 54, 56]. Other SSFL methods design more robust global models [9, 55, 63] to generate pseudo-labels for unlabeled data. However, the global model captures the global data distribution, which is different from clients' local data distribution in non-IID scenarios. In this case, distribution mismatch may lower pseudo-label accuracy, degrading model performance.

One possible solution to handle distribution mismatch is to enhance pseudo-labels with local knowledge. Figure 1 gives a label-skew example in non-IID SSFL scenarios, where some data is randomly sampled from unlabeled data for labeling. We observe that distributions of local labeled data and unlabeled data are consistent and imbalanced due to random sampling. It indicates a local model that captures the local data distribution can generate accurate pseudo-labels for most local unlabeled data. Besides, the global data distribution is highly different from local data distributions. This means pseudo-labels generated by the global model is subpar. We further validate this insight under different non-IID degrees in Figure 2. When the heterogeneity increases, the local models trained on local labeled data alone can achieve better pseudo-label accuracy than the global model trained with the SOTA SSFL method (FedRGD [63]). This result denotes that local knowledge is helpful to accurate pseudo-labeling in non-IID scenarios.

In view of this phenomenon, we propose a novel SSFL technique named **FedLoKe**. The core idea of FedLoKe is exploiting local knowledge to enhance pseudo-labeling in non-IID scenarios. Specifically, we first introduce a global-local co-training structure, which includes a local model for each client to capture the local biased data distribution and a global model shared by all clients to capture the global data distribution. As the local model well captures the local data distribution, we use it to generate better pseudo-labels of the unlabeled dataset for the global model. The global model is then trained on the local labeled dataset and local pseudo-label dataset. Moreover, to mitigate local model overfitting, we update the local model with the global model via Exponential Moving Average (EMA) [22]. Meanwhile, besides the local labeled dataset, the local model is also trained on the global pseudo-label dataset generated by the global model. This transfers general knowledge to the local model and further prevents it from overfitting.

Our contributions are summarized as follows:

- We show the benefits brought by non-IID on pseudo-labeling in SSFL. This inspires us to exploit local knowledge to alleviate the problem of pseudo-label difficulty in heterogeneous SSFL.
- We propose a novel SSFL method called FedLoKe, which leverages local knowledge to enhance pseudo-labeling in non-IID scenarios for global model training and utilizes global knowledge to mitigate the overfitting problem of local model training.
- We empirically validate the effectiveness of FedLoKe on four benchmark image datasets. The result shows FedLoKe can achieve better performance than the state-of-the-art SSFL methods.

2 RELATED WORK

2.1 Federated Learning

FL is a technique which can protect decentralized privacy-sensitive data. Instead of collecting data stored on the client side, the server gathers the local model updates from selected clients and aggregates the updates to form a new model [28, 29, 38]. However, these methods suffer severe performance degradation when the client data is non-IID. Many works try to mitigate the impact of data heterogeneity via building a more robust model [18, 28, 33, 37]. Another solution to the data heterogeneity is personalized federated learning (pFL) [8, 29, 41, 47]. Besides training a global model, pFL also learns a personalized model for each client to mitigate the effect of negative transfer caused by non-IID. The core idea of pFL is similar with FedLoKe, as both utilize local knowledge. However, the objectives of FedLoKe and pFL differ fundamentally. pFL focuses on leveraging local models to improve generalization on each client's local data [8, 29]. In contrast, the primary goal of FedLoKe's local models is to assist the global model to learn more accurate knowledge from unlabeled data. Furthermore, similar to other FL methods, pFL operates under the assumption that all data is labeled. In more practical scenarios, there are few data is annotated and most of the data is unlabeled. The presence of unlabeled data renders pFL unsuitable in this scenario.

2.2 Semi-Supervised Learning

Annotating data is expensive and time-consuming. Hence, semi-supervised learning (SSL) utilizes limited labeled data and abundant unlabeled data for model training, yielding performance comparable to supervised learning. There are two popular techniques in SSL, i.e., consistency regularization and pseudo-labeling. Consistency regularization [1, 46] encourages models to generate the same predictions when the inputs are perturbe, while pseudo-labeling [23] views confident predictions produced by models as the hard [42, 50] or soft [2, 57] labels of the unlabeled data and add pseudo-labeled data to training. A number of recent works integrate both techniques [2, 25, 50, 61]. Specifically, FixMatch [50] uses hard pseudo labels on weakly-augmented unlabeled data to guide the predictions of strongly-augmented unlabeled data. However, these algorithms are based on central settings. When the data are distributed to different clients, the more limited labeled data on each client and the different client distributions cause more complex problems. For example, in FL scenarios, it is challenging to satisfy the class match assumption [10, 11, 13, 32, 53, 55, 64], which assumes that unlabeled data has classes that only appear in labeled data.

2.3 Semi-Supervised Federated Learning

Recently, SSFL has gained significant prominence [3, 9, 15, 17, 36, 63]. Researchers in this domain mainly focus on three SSFL scenarios. The first two scenarios suppose that labeled data is solely accessible on the central server [3, 9, 15, 17, 36, 63] or on some clients [26, 31, 34, 58], and remaining clients only own unlabeled data. The third scenario assumes that each client has limited labeled data and massive unlabeled data. We concentrate on the third scenario. Many works also consider this scenario [15, 20, 32, 49, 55, 63]. They try to alleviate the negative impact of non-IID in SSFL in two ways. The first solution is to ensemble models from other clients to generate more reliable pseudo-labels [15, 20, 32, 49]. Another method proposes a more robust aggregation mechanism to reduce the impact of erroneous pseudo-labels on the global model [55, 63]. However, all these methods disregard or underestimate the benefits of local knowledge brought by non-IID for pseudo-labeling in heterogeneous SSFL. Unlike these works, we exploit local knowledge to help the global model learn from unlabeled data.

3 PROBLEM DEFINITION

Formally, the system has N clients and a server. Each client $i \in \{1, ..., N\}$ has a local model Θ_i^l to capture the local biased distribution and its own private training datasets consist of labeled set $\mathcal{D}_i^l$ and unlabeled set $\mathcal{D}_i^u$. We use (x_s^l, y_s^l) to denote the s-th instance in $\mathcal{D}_i^l$, and use x_s^u to denote the s-th instance in $\mathcal{D}_i^u$. All clients have the same labeling ratio $r = |\mathcal{D}_i^l|/(|\mathcal{D}_i^l| + |\mathcal{D}_i^u|)$, which is often small. In the server, there is a global model Θ^g shared by all clients. Our goal is to effectively learn a global model Θ^g.

4 METHODOLOGY

In this section, we provide an introduction to FedLoKe. We first give an overview of FedLoKe. Then we show the details of the local model updating, which is the core innovation of our method.

4.1 Overview

As shown in Figure 3 (a), similar to existing FL methods [38], our method contains three steps in each training round t, i.e., model distribution, model updating and model aggregation. First, the server randomly samples a group of clients C_t and distributes the global model Θ_t^g to them. Second, each sampled client i trains the local model $\Theta_{i,t}^l$ and the received global model $\Theta_{i,t}^g$ on both the local labeled data $\mathcal{D}_i^l$ and unlabeled data $\mathcal{D}_i^u$. Finally, the server aggregates the trained global models $\Theta_{i,t+1}^g$ of the sampled clients as follows:

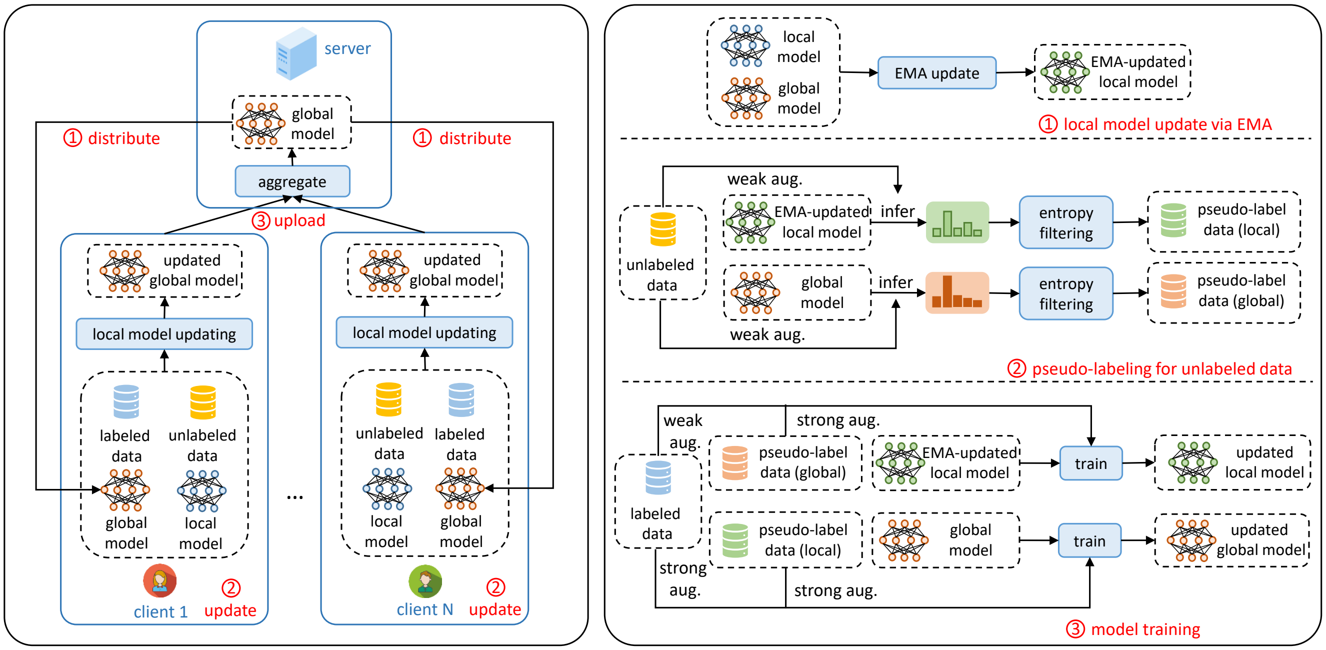
$$\Theta_{t+1}^g = \frac{\sum_{i \in C_t} (|\mathcal{D}_i^l| + |\mathcal{D}_i^u|)\Theta_{i,t+1}^g}{\sum_{j \in C_t} (|\mathcal{D}_j^l| + |\mathcal{D}_j^u|)}. \tag{1}$$

The above steps repeat T rounds until the global model converges.

The key innovation of FedLoKe is the model updating step. The details are shown in Figure 3 (b), which consists of three steps for each selected client i in the t-th round, i.e., local model update, pseudo-labeling and model training. In the local model update, we update the local model $\Theta_{i,t}^l$ with the received global model $\Theta_{i,t}^g$ via EMA. In the pseudo-labeling, the local and global models generate pseudo-labels for unlabeled data, creating the local and global pseudo-label datasets, respectively. In the model training, the local model trains on the labeled data and global pseudo-label dataset, while the global model trains on the labeled data and local pseudo-label dataset. The training process is shown in Algorithm1. We will provide the details of the three steps in following subsections.

4.2 Local Model Update via EMA

Since local models only learn from limited local data, they easily overfit and get stuck in local minima. At the same time, local models also drift far from the global model in non-IID conditions, which



(a) overview of FedLoke

(b) details of the model updating step of FedLoke

Figure 3: The overall framework of FedLoKe.

may generate severely biased pseudo-labels [19, 35]. To incorporate global knowledge beyond local data into local models and decrease the distance between local and global models, inspired by [52], local models are fused with the global model before training with data:

$$\Theta_{i,t}^{l'} = \mu\Theta_{i,t}^{l} + (1-\mu)\Theta_{i,t}^{g}, \tag{2}$$

where μ is a hyperparameter which controls the proportion of global model and local model fusion.

4.3 Pseudo-Labeling for Unlabeled Data

To expand the training data and provide more training signals, we utilize pseudo-labeling for unlabeled data. The pseudo-labeling contains two steps, i.e., inferring soft labels and uncertainty filtering.

The first step is inferring soft labels. We feed weakly augmented samples into the local and global model to generate reliable pseudo-labels for the unlabeled data:

$$\hat{\boldsymbol{y}}_s^l = f(\alpha(x_s^u); \Theta_{i,t}^{l'}), \tag{3}$$

$$\hat{\boldsymbol{y}}_s^g = f(\alpha(x_s^u); \Theta_{i,t}^{g}), \tag{4}$$

where $\hat{\boldsymbol{y}}_s^l$ and $\hat{\boldsymbol{y}}_s^g$ are the classification prediction probability vectors of all classes, $f(\cdot)$ is model architecture function and $\alpha(\cdot)$ is weak data augmentation function which uses a standard flip-and-shift augmentation strategy [50].

The second step is uncertainty filtering. Since the model predictions are noisy and possibly wrong, it is important to filter the incorrect pseudo-labels to reduce the error rate. A commonly applied method is to filter out the samples with high-uncertainties. Existing works have studied using the highest confidence of predictions [32, 50], performing label propagation on the nearest feature space [48] and applying the entropy of predictions [40]. In our FedLoKe, we filter them based on the entropy of predictions:

$$E_s^l = -\sum_{j=1}^{C} \hat{\boldsymbol{y}}_s^l[j] log(\hat{\boldsymbol{y}}_s^l[j]), \tag{5}$$

$$E_s^g = -\sum_{j=1}^{C} \hat{\boldsymbol{y}}_s^g[j] log(\hat{\boldsymbol{y}}_s^g[j]), \tag{6}$$

where C is the number of classes, $\hat{\boldsymbol{y}}_s^l[j]$ is the probability for the j-th class of the local model, $\hat{\boldsymbol{y}}_s^g[j]$ is the probability for the j-th class of the global model. E_s^l and E_s^g correspond to the prediction entropies of the local and global models, respectively. The lower entropy means the higher certainty of the model prediction. We set a hyperparameter threshold δ that controls how certainly a prediction can be used as pseudo-labels. We filter the uncertain samples to form pseudo-label datasets:

$$\mathcal{D}_i^{pl} = \{(x_s^u, \hat{\boldsymbol{y}}_s^l) | x_s^u \in \mathcal{D}_i^u, E_s^l < \delta\}, \tag{7}$$

$$\mathcal{D}_i^{pg} = \{(x_s^u, \hat{\boldsymbol{y}}_s^g) | x_s^u \in \mathcal{D}_i^u, E_s^g < \delta\}, \tag{8}$$

where $\mathcal{D}_i^{pl}$ and $\mathcal{D}_i^{pg}$ represent the local and global pseudo-label dataset of the i-th client, respectively.

4.4 Model Training

In this subsection, we introduce the model training step, which consists of three sub-steps, i.e., pseudo-label data learning, labeled data learning and model updating.

The first step is pseudo-label data learning. Since the amount of labeled data in clients' devices is small in SSFL, it is crucial for

Algorithm 1: FedLoKe

```
Server initializes $\Theta_0^g$ randomly;
Each client $i$ initializes $\Theta_{i,0}^l$ randomly;
for each round $t = 0, ..., T$ do
    Randomly select a client set $C_t$ from $N$ clients;
    Distribute the global model $\Theta_t^g$ to selected clients;
    for client $i \in C_t$ do
        $\Theta_{i,t+1}^g \leftarrow$ ClientUpdate$(i, \Theta_t^g)$;
        Upload global model $\Theta_{i,t+1}^g$ to server;
    end
    Aggregate updated global models to obtain $\Theta_{t+1}^g$ (Eq.1);
end
ClientUpdate:
    $\Theta_{i,t}^{l'} \leftarrow$ EMA$(\Theta_{i,t}^l, \Theta_{i,t}^g)$ (Eq.2) ;
    $\mathcal{B}_u \leftarrow$ (split $\mathcal{D}_i^u$ into batches of size $B$) ;
    $\mathcal{B}_l \leftarrow$ (split $\mathcal{D}_i^l$ into batches of size $B$) ;
    for each unlabeled batch data $x_s^u \in \mathcal{B}_u$ do
        Select a labeled batch data $(x_k^l, y_k^l) \in \mathcal{B}_l$ ;
        Generate two pseudo-label datasets for $x_s^u$ (Eq.7,8);
        Compute the loss for the global model (Eq.16);
        Compute the loss for the local model (Eq.15);
        Update local model $\Theta_{i,t}^{l'}$ and global model $\Theta_{i,t}^g$ ;
    end
    $\Theta_{i,t+1}^g \leftarrow \Theta_{i,t}^g$, $\Theta_{i,t+1}^l \leftarrow \Theta_{i,t}^{l'}$ ;
    return $\Theta_{i,t+1}^g$;
```

models to learn on the pseudo-label dataset. As shown in Figure 1, the data distributions of labeled and unlabeled data are consistent and imbalanced in non-IID SSFL scenarios, while the global data distribution is different from the local data distribution. Besides, as shown in Figure 2, the pseudo-label accuracy of local models is higher than the pseudo-label accuracy of global model, especially in highly non-IID settings. It indicates local knowledge is helpful in generating accurate pseudo-labels for most local unlabeled data in non-IID scenarios. Therefore, in FedLoKe, we propose to train the global model on the local pseudo-label dataset $\mathcal{D}_i^{pl}$. Meanwhile, if the local model only learns on the local labeled dataset or teaches itself using the local pseudo-label dataset, it is easily overfitted [4]. We mitigate the overfitting problem by transfer the general global knowledge to the local model. Therefore, we train the local model on the global pseudo-label dataset $\mathcal{D}_i^{pg}$.

To teach both models more general understanding, we leverage strongly augmented versions of the same images in pseudo-label datasets as the more diverse input of models:

$$\widetilde{\mathcal{L}}_i^{U,g} = \frac{1}{|\mathcal{D}_i^u|} \sum_{(x_s^u, \hat{\boldsymbol{y}}_s^l) \in \mathcal{D}_i^{pl}} KL(\hat{\boldsymbol{y}}_s^l, f(\mathcal{A}(x_s^u); \Theta_{i,t}^g)), \tag{9}$$

$$\widetilde{\mathcal{L}}_i^{U,l} = \frac{1}{|\mathcal{D}_i^u|} \sum_{(x_s^u, \hat{\boldsymbol{y}}_s^g) \in \mathcal{D}_i^{pg}} KL(\hat{\boldsymbol{y}}_s^g, f(\mathcal{A}(x_s^u); \Theta_{i,t}^{l'})), \tag{10}$$

where $KL(\cdot)$ is the Kullback–Leibler divergence function and $\mathcal{A}(\cdot)$ is the strong data augmentation function using RandAugment [6]. $\widetilde{\mathcal{L}}_i^{U,g}$ and $\widetilde{\mathcal{L}}_i^{U,l}$ denote the global and local distillation losses, respectively. We average the two losses using the number of all unlabeled data. This is because every unlabeled data should carry equal weight regardless of the pseudo-label dataset's size. Since the initial local and global model are not stable and may generate low-quality pseudo-label dataset, we set the weight of unlabeled loss small at the early rounds and gradually increase the weight as follows:

$$\mathcal{L}_i^{U,l} = min(1, \frac{t}{T_u})\widetilde{\mathcal{L}}_i^{U,l}, \tag{11}$$

$$\mathcal{L}_i^{U,g} = min(1, \frac{t}{T_u})\widetilde{\mathcal{L}}_i^{U,g}, \tag{12}$$

where T_u is a hyperparameter that regulates the intensity of the unlabeled loss.

The second step is labeled data learning. We simultaneously train local and global models on labeled data to safeguard against forgetting labeled data [15] and alleviate the impacts of erroneous pseudo-labels causing misguidance. To make local models learn more concrete and reliable local knowledge helping pseudo-labeling, every client trains its local model $\Theta_{i,t}^{l'}$ on the weakly-augmented labeled data using cross-entropy loss:

$$\mathcal{L}_i^{L,l} = \frac{1}{|\mathcal{D}_i^l|} \sum_{(x_s^l, y_s^l) \in \mathcal{D}_i^l} CE(y_s^l, f(\alpha(x_s^l); \Theta_{i,t}^{l'})), \tag{13}$$

where $CE(\cdot)$ means the cross entropy loss function. Since applying the strong augmentation $\mathcal{A}$ on labeled data can make models learn more general knowledge [6, 44], we feed the strongly-augmented labeled data to the global model:

$$\mathcal{L}_i^{L,g} = \frac{1}{|\mathcal{D}_i^l|} \sum_{(x_s^l, y_s^l) \in \mathcal{D}_i^l} CE(y_s^l, f(\mathcal{A}(x_s^l); \Theta_{i,t}^g)). \tag{14}$$

The final step is model updating. The overall loss of the local model is:

$$\mathcal{L}_i^l = \mathcal{L}_i^{L,l} + \mathcal{L}_i^{U,l}. \tag{15}$$

The overall loss of the global model is:

$$\mathcal{L}_i^g = \mathcal{L}_i^{L,g} + \mathcal{L}_i^{U,g}. \tag{16}$$

In each local batch, we update both models with learning rate η. After training $|\mathcal{D}_i^u|/B$ batches, we get the updated global model $\Theta_{i,t+1}^g$ and the update local model $\Theta_{i,t+1}^l$, where B is the batch size.

5 EXPERIMENT

In this section, we validate the effectiveness of FedLoKe.

5.1 Experimental Settings

Datasets. Following existing works in SSFL [9, 20, 63], we use four benchmark image datasets for validation, including CIFAR-10 [21], SVHN [39], CIFAR-100 [21] and TinyImageNet [7]. The detailed statistics are shown in Table 3. Following previous works [3, 9, 32], we use the Dirichlet distribution $Dir(\alpha)$ to simulate non-IID data distribution. The smaller α indicates a more imbalanced distribution. We utilize the $Dir(\alpha)$ distribution to sample an equivalent number of training images for each client from the remaining training images. To simulate a semi-supervised environment whose label ratio is r, we randomly pick r images from every client dataset as labeled data, and the remaining images are unlabeled data.

Table 1: Performance of different methods under different label ratios on CIFAR-10 and SVHN.

	CIFAR-10			SVHN		
	$r = 1\%$	$r = 5\%$	$r = 10\%$	$r = 1\%$	$r = 5\%$	$r = 10\%$
FedAvg-SL	86.67 ± 0.32	86.67 ± 0.32	86.67 ± 0.32	94.12 ± 0.11	94.12 ± 0.11	94.12 ± 0.11
FixMatch	64.02 ± 2.30	85.68 ± 0.40	89.17 ± 0.20	92.99 ± 0.37	95.63 ± 0.28	96.16 ± 0.17
FedAvg	38.69 ± 2.16	55.54 ± 0.71	64.60 ± 0.89	38.31 ± 1.90	84.84 ± 0.29	86.18 ± 0.39
FixMatch-FedAvg	39.84 ± 2.20	55.91 ± 1.68	59.45 ± 1.51	49.15 ± 7.49	92.43 ± 0.38	93.34 ± 0.35
FML	32.74 ± 1.65	51.95 ± 0.39	61.89 ± 0.38	21.28 ± 0.67	44.00 ± 1.65	74.80 ± 0.78
FedMatch	37.60 ± 0.68	48.04 ± 0.21	53.69 ± 0.56	45.51 ± 2.69	79.24 ± 0.44	86.63 ± 0.06
semiFed	21.09 ± 2.73	52.37 ± 1.90	55.16 ± 0.76	25.02 ± 1.69	64.11 ± 3.15	88.21 ± 0.55
FedRGD	44.84 ± 1.80	61.47 ± 1.48	68.78 ± 1.56	83.69 ± 7.89	93.87 ± 0.09	94.39 ± 0.29
FedLoKe	**53.14 ± 3.53**	**77.94 ± 0.76**	**81.99 ± 0.30**	**93.52 ± 0.33**	**94.76 ± 0.15**	**95.13 ± 0.12**

Table 2: Performance of different methods under different label ratios on CIFAR-100 and TinyImageNet.

	CIFAR-100			TinyImageNet		
	$r = 1\%$	$r = 5\%$	$r = 10\%$	$r = 1\%$	$r = 5\%$	$r = 10\%$
FedAvg-SL	58.25 ± 0.38	58.25 ± 0.38	58.25 ± 0.38	39.59 ± 0.57	39.59 ± 0.57	39.59 ± 0.57
FixMatch	13.37 ± 0.83	38.59 ± 0.90	52.01 ± 1.12	6.93 ± 0.32	22.60 ± 0.48	31.93 ± 0.44
FedAvg	5.56 ± 0.73	17.38 ± 0.10	26.50 ± 0.43	3.12 ± 0.24	8.83 ± 0.20	14.04 ± 0.23
FixMatch-FedAvg	10.35 ± 0.77	24.74 ± 0.29	30.96 ± 0.56	5.38 ± 0.44	13.93 ± 0.28	19.88 ± 0.26
FML	8.09 ± 0.42	19.76 ± 0.24	27.81 ± 0.44	4.50 ± 0.16	11.51 ± 0.24	16.85 ± 0.39
FedMatch	7.24 ± 0.58	14.13 ± 0.45	18.57 ± 0.33	2.65 ± 0.28	4.74 ± 0.13	6.41 ± 0.39
semiFed	5.67 ± 0.46	14.60 ± 0.31	23.83 ± 0.27	1.57 ± 0.34	6.65 ± 0.21	11.85 ± 0.25
FedRGD	11.09 ± 0.61	22.06 ± 0.60	28.66 ± 0.71	5.30 ± 0.47	11.13 ± 0.09	15.26 ± 0.21
FedLoKe	**13.23 ± 0.59**	**33.83 ± 0.27**	**44.49 ± 0.62**	**7.27 ± 0.35**	**19.53 ± 0.26**	**26.54 ± 0.33**

Table 3: Detailed statistics of dataset.

	CIFAR-10	SVHN	CIFAR-100	TinyImageNet
# training samples	50,000	73,257	50,000	100,000
# test samples	10,000	26,032	10,000	10,000
# classes	10	10	100	200

Evaluation task. Following [9, 15, 36], we evaluate the global model using the accuracy of the global model on the test dataset. For robustness, we replicate each experiment 5 times independently and present the mean performance along with standard deviations.

Implementation details. We assume there are 50 clients and select 20 clients randomly to participate in each round of training. We set δ to 0.1, T_u to 200, η to 0.01, local batch size to 32, local epoch to 1 and training rounds to 500. As μ relies on various factors (e.g. dataset and heterogeneity), we tune $\mu \in \{0.5, 0.7, 0.9\}$. Following [15, 20], we utilize ResNet9 [12] as the backbone model.

5.2 Performance Evaluation

We compare our FedLoKe with the following methods.

- **FedAvg** [38], which utilizes only labeled datasets $\mathcal{D}_i^l$ in every clients to train models. This is a lower bound for SSFL methods which does not consider unlabeled dataset.
- **FixMatch-FedAvg** [15], applying FixMatch as the local model training algorithm in conjunction with FedAvg.
- **FML** [47], which is a pFL method. Local models and global models in FML use cross entropy and mutual knowledge distillation [62] to learn labeled data. For a fair comparison, FML also leverages mutual knowledge distillation on unlabeled data.
- **FedMatch** [15], which incorporates inter-client consistency loss and parameter decomposition for disjoint learning on labeled and unlabeled data.
- **semiFed** [32], which adopts inter-client helpers to generate pseudo labels.
- **FedRGD** [63], utilizing group normalization and group aggregation to generate more robust global models in SSFL.
- **FedAvg-SL** [15], which uses all data in every clients as labeled data to conduct supervised learning. This serves as an upper bound of SSFL methods because all data are labeled data.
- **FixMatch** [50], central FixMatch applied to the central semi-supervised dataset gathered from all clients' labeled and unlabeled data. This serves as an upper bound of SSFL methods because it doesn't encounter issues of distributed data.

Following [31, 32], we conduct three experiments with $r \in \{1\%, 5\%, 10\%\}$ labeled data ratio under $Dir(0.5)$. We show the results of all methods in Table 1 and 2.

We have several findings. First, the performance of FedLoKe is significantly higher than the lower bound provided by FedAvg. This is because FedLoKe effectively extracting knowledge from unlabeled data using pseudo-labeling. Secondly, FedLoKe outperforms FixMatch-FedAvg sharply. This indicates that the straightforward combination of SSL and FL is not adequate to achieve a satisfactory performance in SSFL. Thirdly, the performance of FedLoKe is better than FML. This stems from the fact that pFL mainly focuses on supervised learning, thereby lacking specific designs to effectively tackle the challenges in SSFL. Fourthly, FedLoKe performs better

Table 4: Performance of different methods under different non-IID degrees on CIFAR-10 and SVHN.

	CIFAR-10			SVHN		
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.5$
FedAvg-SL	55.84 ± 5.15	81.20 ± 0.87	86.67 ± 0.32	87.33 ± 1.14	92.83 ± 0.30	94.12 ± 0.11
FixMatch	85.68 ± 0.40	85.68 ± 0.40	85.68 ± 0.40	95.63 ± 0.28	95.63 ± 0.28	95.63 ± 0.28
FedAvg	25.59 ± 5.84	45.31 ± 2.59	55.54 ± 0.71	80.76 ± 1.42	82.70 ± 0.77	84.84 ± 0.29
FixMatch-FedAvg	28.72 ± 5.05	44.49 ± 3.28	55.91 ± 1.68	83.69 ± 1.56	90.90 ± 0.55	92.43 ± 0.38
FML	32.24 ± 4.24	46.05 ± 2.83	51.95 ± 0.39	33.97 ± 2.50	42.58 ± 2.05	44.00 ± 1.65
FedMatch	45.26 ± 0.56	46.05 ± 0.54	48.04 ± 0.21	77.28 ± 1.55	76.91 ± 0.83	79.24 ± 0.44
semiFed	32.03 ± 3.11	47.27 ± 0.70	52.37 ± 1.90	54.70 ± 4.73	61.12 ± 2.43	64.11 ± 3.15
FedRGD	34.47 ± 2.79	53.10 ± 1.88	61.47 ± 1.48	**87.01 ± 2.22**	92.62 ± 0.68	93.87 ± 0.09
FedLoKe	**48.51 ± 3.41**	**70.30 ± 1.10**	**77.94 ± 0.76**	86.84 ± 1.45	**93.77 ± 0.29**	**94.76 ± 0.15**

Table 5: Performance of different methods under different non-IID degrees on CIFAR-100 and TinyImageNet.

	CIFAR-100			TinyImageNet		
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.5$
FedAvg-SL	40.27 ± 0.94	55.08 ± 0.39	58.25 ± 0.38	27.98 ± 0.97	37.50 ± 0.34	39.59 ± 0.57
FixMatch	38.59 ± 0.90	38.59 ± 0.90	38.59 ± 0.90	22.60 ± 0.48	22.60 ± 0.48	22.60 ± 0.48
FedAvg	13.33 ± 0.58	16.92 ± 0.43	17.38 ± 0.10	7.01 ± 0.65	8.98 ± 0.31	8.83 ± 0.20
FixMatch-FedAvg	12.27 ± 1.85	22.36 ± 0.81	24.74 ± 0.29	8.16 ± 0.57	12.34 ± 0.51	13.93 ± 0.28
FML	13.45 ± 0.48	18.47 ± 0.58	19.76 ± 0.24	7.13 ± 0.38	10.76 ± 0.32	11.51 ± 0.24
FedMatch	12.19 ± 0.33	13.28 ± 0.20	14.13 ± 0.45	3.82 ± 0.24	4.55 ± 0.26	4.74 ± 0.13
semiFed	11.07 ± 0.45	14.14 ± 0.36	14.60 ± 0.31	5.14 ± 0.48	6.86 ± 0.28	6.65 ± 0.21
FedRGD	16.80 ± 0.59	20.81 ± 0.33	22.06 ± 0.60	9.40 ± 0.44	10.48 ± 0.35	11.13 ± 0.09
FedLoKe	**28.45 ± 1.32**	**34.86 ± 0.85**	**33.83 ± 0.27**	**16.26 ± 0.77**	**19.92 ± 0.18**	**19.53 ± 0.26**

than FedMatch and semiFed, which utilize other clients’ models to help pseudo-labeling. This highlights the difficulty in selecting appropriate helper clients for each client due to heterogeneity. Inconsiderate selection of helper clients will cause serious misleading on unlabeled data. FedLoKe exclusively leverages local knowledge for pseudo-labeling, which has been showed to be more reliable in non-IID scenarios, as illustrated in Figure 2. Fifthly, FedLoKe achieves better performance than FedRGD, which aims to obtain a more robust global model. This denotes that the higher quality of pseudo-labels brought by local knowledge benefits the global model significantly. Last, our FedLoKe narrows the gap with both upper bounds, i.e., FedAvg-SL and FixMatch. This shows the effectiveness of incorporating local knowledge enhancement in SSFL.

5.3 Influence of non-IID Degrees

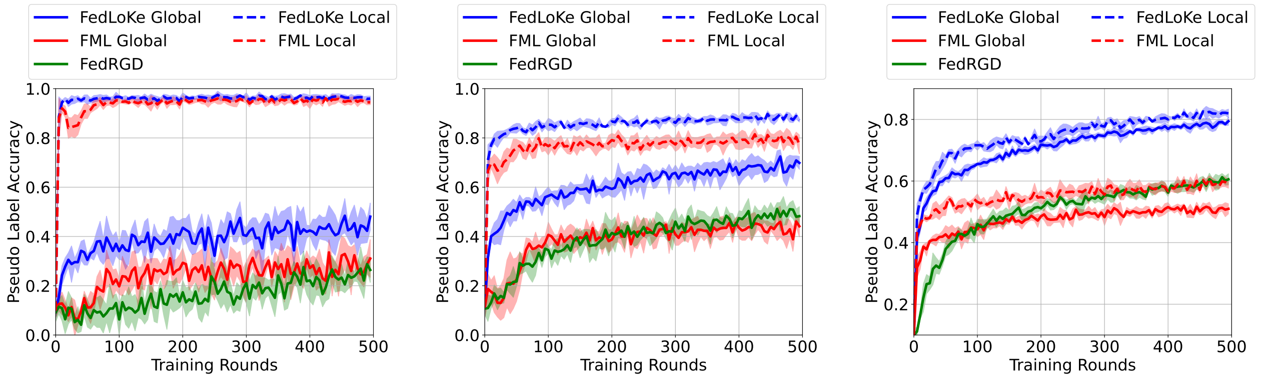
In this subsection, we first explore the influence of non-IID degrees on the performance of FedLoKe and baseline methods. Then, we discuss the impact of non-IID degrees to pseudo-label accuracy.

We evaluate SSFL methods under three different non-IID distributions, i.e., $\alpha \in \{0.01, 0.1, 0.5\}$, which all have label ratios $r = 5\%$. The results are shown in Table 4 and 5. We have several observations. First, all methods usually achieve better performance when non-IID degree is weaker. This is because the greater the degree of non-IID, the more difficult it is for the server to obtain a optimal global model by aggregating divergent client models. Second, FedLoKe outperforms other methods in most situations. This is because FedLoKe can leverage local knowledge to enhance pseudo-labeling in non-IID scenarios. Last, the higher heterogeneous environment can improve the accuracy of pseudo-labeling as illustrated in Figure 2, but still hurts FedLoKe performance when distributions are extremely non-IID. This is because FedLoKe still faces problems caused by non-IID, such as client drifting [16, 18], even though most pseudo-labels are correct. We plan to overcome the impact of non-IID to model learning while maintaining the advantages of pseudo-label brought by heterogeneity as future works.

Next, to explore the impact of non-IID degree to pseudo-label accuracy of FedLoKe, we show the pseudo-label accuracy of FedLoKe, FML and FedRGD under different non-IID degrees in Figure 6. We have several findings. First, we observe that local models of FedLoKe consistently achieve better pseudo-label accuracy compared to the global model of FedLoKe in heterogeneous environments. This observation is in line with our expectation as shown in Figure 2. Therefore, the pseudo-label generated by local models can help the global model learn better from the unlabeled data. Furthermore, the performance of local models improves as the level of heterogeneity increases. This reinforces the idea that non-IID is not always bad for SSFL, because local models are more effective at pseudo-labeling in more heterogeneous environments. Besides, FedRGD consistently underperforms the global model of FedLoKe, which shows the advantages of leveraging local knowledge for SSFL. Lastly, local and global models of FedLoKe outperform local and global models of FML, showcasing our specific designs for SSFL is beneficial.

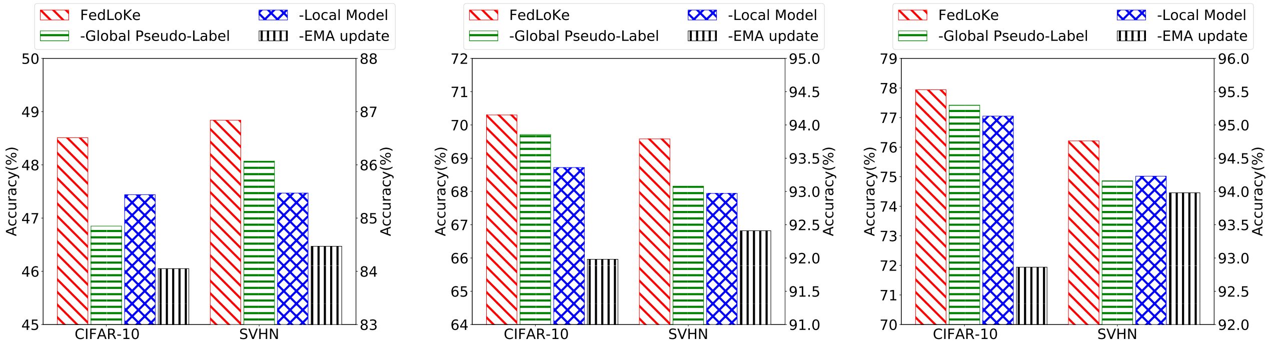
5.4 Ablation Study

In this subsection, we first explore the effectiveness of three key designs, i.e., Local Models, Global Pseudo-Label, and EMA Update. And then we validate the role of data augmentation in FedLoKe.



(a) α = 0.01 (b) α = 0.1 (c) α = 0.5

Figure 4: Pseudo-label accuracy of different models under different non-IID degrees on CIFAR-10.



(a) α = 0.01 (b) α = 0.1 (c) α = 0.5

Figure 5: Ablation Study of FedLoKe on CIFAR-10 and SVHN under different non-IID degrees.

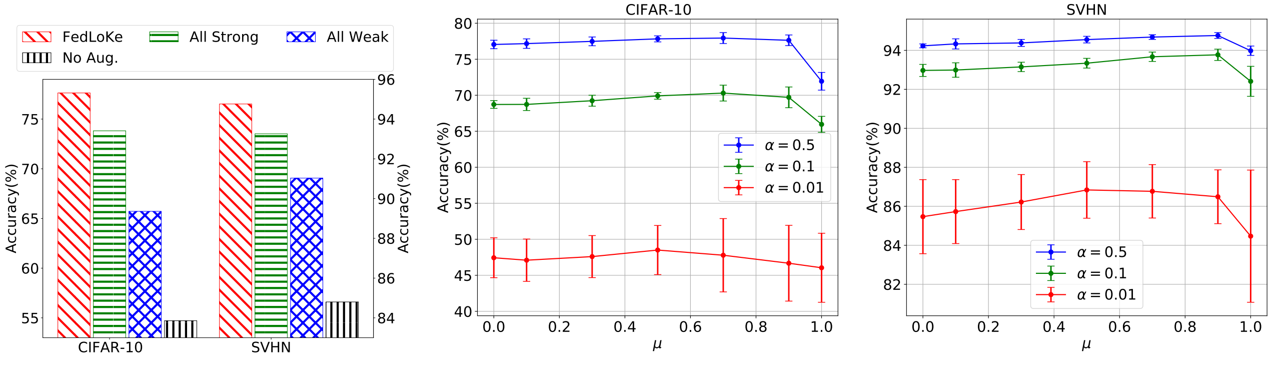
We show the results of CIFAR-10 and SVHN under different non-IID degrees in Figure 5. We have several observations. First, replacing local models with the global model hurts the performance. This is because local knowledge from local models is valuable for pseudo-labeling in non-IID conditions. Second, the performance of the global model without generating a pseudo-label (Global Pseudo-Label) to teach the local model on unlabeled data also drops. This highlights the importance of Global Pseudo-Label transferring global knowledge to the local model. Global Pseudo-Label can mitigate the overfitting problem of local model, and assist the local model in learning the minority classes of local unlabeled data, which is hard to be pseudo-labeled correctly only based on biased local distributions [60]. Last, removing the EMA update reduces the model performance. This means the EMA update can prevent the local model from overfitting on the small number of local data, resulting in higher-quality pseudo-labels.

Next, to validate the augmentation strategy used in FedLoKe is effective, we propose three variants: **All Strong** replaces all data augmentation with RandAugment [6]. **All Weak** replaces all data augmentation with a standard flip-and-shift augmentation [50]. **No Aug.** inputs the original images directly. The results are shown in Figure 6(a). We have several findings. First, we find that our augmentation strategy performs better than All Strong. This indicates that models may generate more unreliable pseudo-labels due to the more diverse augmented input. Second, the performance of All Weak also decreases. This is because All Weak lacks enough diverse input, and violates consistency regularization [50]. Last, the performance of No Aug. drops sharply. This is because models overfit on specific images easily without data augmentation.

5.5 Influence of Unlabeled Data Amount

In this subsection, we explore the influence of unlabeled data amount on FedLoKe and FedGRD under different ratios of used unlabeled data. We randomly select a certain proportion of unlabeled data from the local unlabeled dataset for training. The label ratio is set as 5% and α is set as 0.5 in this experiment.

The results are shown in Figure 7, where we have multiple observations. First, FedLoKe can improve the performance with more unlabeled data. This shows that FedLoKe can extract more potential knowledge from more unlabeled data. Second, the results also



(a)

(b)

(c)

Figure 6: Figure 6(a) shows performance of FedLoKe using different augmentation strategy. Figure 6(b) and 6(c) show performance of FedLoKe with different EMA parameters μ on CIFAR-10 and SVHN under different non-IID degrees α.

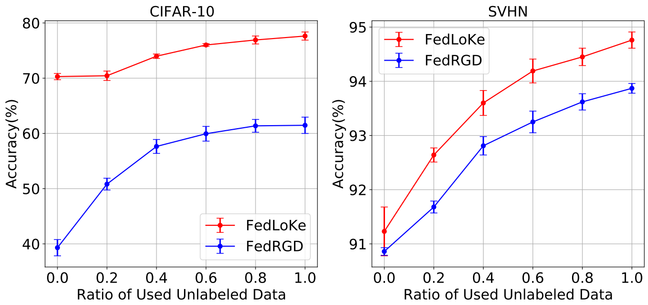


Figure 7: Performance of FedLoKe and FedRGD under different ratios of used unlabeled data.

Table 6: Performance of different methods under different client numbers on CIFAR-100.

Client Number	FedRGD	FedLoKe
30	22.96 ± 0.54	36.17 ± 0.75
40	22.42 ± 0.23	34.93 ± 0.88
50	22.06 ± 0.60	33.83 ± 0.27
60	21.45 ± 0.56	32.47 ± 0.27

reveal that FedLoKe can utilize unlabeled data more effectively than FedRGD under different ratios of used unlabeled data. This is because FedLoKe incorporates local knowledge to enhance the accuracy of pseudo-labeling, while FedRGD only relies on the global model to generate pseudo-labels, which may deviate from the local distributions resulting in sub-optimal performance.

5.6 Influence of EMA weight

In this subsection, we explore how the EMA weight μ influences the performance of FedLoKe. We show these results in Figure 6(b) and 6(c). We notice a performance enhancement as μ rises. This is because small μ makes local models difficult to remember local distributions, hindering mining local knowledge from unlabeled data. However, when μ is too large, the performance drops sharply. This is because local models easily overfit on imbalanced data and stuck in local minima without the global model updates [60], resulting in the global model to be misled by local models.

5.7 Influence of Client Number

We conduct experiments on CIFAR-100 for FedLoKe and FedRGD under different client numbers, as shown in Table 6. We find that FedLoKe outperforms FedRGD consistently.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel semi-supervised heterogeneous federated learning method, named FedLoKe, which focuses on generating accurate pseudo-labels with local knowledge to boost the performance of the global model. In FedLoKe, we create a local model for each client to capture the local data distribution and a global model shared by all clients to capture the global data distribution. To effectively train the global model, we propose to use the local model to generate accurate pseudo-labels for the most unlabeled dataset, and train the global model with both the labeled dataset and pseudo-labeled dataset. To avoid local models overfitting on sparse local labeled data, we distill the general knowledge of the global model on the unlabeled dataset to the local model. Besides, we use EMA to update the local model with the global model. Experiments show the effectiveness of FedLoKe.

The benefits of local knowledge brought by non-IID on pseudo-labeling in heterogeneous SSFL is the key insight in this paper. In the future, we will devise a computationally efficient framework with theoretical guarantees that considers local knowledge, and combine it with other SSFL techniques, such as robust aggregation.

ACKNOWLEDGMENTS

This work was supported by the grants from National Natural Science Foundation of China (No.62222213, 62072423), and the USTC Research Funds of the Double First-Class Initiative (No.YD2150002009).

REFERENCES

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. 2014. Learning with pseudo-ensembles. *NIPS* 27 (2014).

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *NIPS* 32 (2019).

[3] Jieming Bian, Zhu Fu, and Jie Xu. 2021. FedSEAL: Semi-Supervised Federated Learning with Self-Ensemble Learning and Negative Learning. *arXiv preprint arXiv:2110.07829* (2021).

[4] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. 2022. Debiased Self-Training for Semi-Supervised Learning. In *NIPS*.

[5] Mingqing Chen, Rajiv Mathews, Tom Ouyang, and Françoise Beaufays. 2019. Federated learning of out-of-vocabulary words. *arXiv preprint arXiv:1903.10635* (2019).

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *NIPS*. 702–703.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. 248–255.

[8] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. 2020. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461* (2020).

[9] Enmao Diao, Jie Ding, and Vahid Tarokh. 2021. SemiFL: Communication efficient semi-supervised federated learning with unlabeled clients. *arXiv preprint arXiv:2106.01432* (2021).

[10] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. 2020. Safe deep semi-supervised learning for unseen-class unlabeled data. In *ICML*. 3897–3906.

[11] Lu Han, Han-Jia Ye, and De-Chuan Zhan. 2023. On Pseudo-Labeling for Class-Mismatch Semi-Supervised Learning. *arXiv preprint arXiv:2301.06010* (2023).

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[13] Rundong He, Zhongyi Han, Yang Yang, and Yilong Yin. 2022. Not all parameters should be treated equally: Deep safe semi-supervised learning under class distribution mismatch. In *AAAI*, Vol. 36. 6874–6883.

[14] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. 2019. The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law* 28, 1 (2019), 65–98.

[15] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. 2021. Federated Semi-supervised Learning with Inter-client Consistency & Disjoint Learning. In *ICLR*.

[16] Meirui Jiang, Zirui Wang, and Qi Dou. 2022. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *AAAI*, Vol. 36. 1087–1095.

[17] Meirui Jiang, Hongzheng Yang, Xiaoxiao Li, Quande Liu, Pheng-Ann Heng, and Qi Dou. 2022. Dynamic Bank Learning for Semi-supervised Federated Image Diagnosis with Class Imbalance. In *MICCAI*. Springer, 196–206.

[18] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*. 5132–5143.

[19] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. 2020. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *NIPS* 33 (2020), 14567–14579.

[20] Woojung Kim, Keondo Park, Kihyuk Sohn, Raphael Shu, and Hyung-Sin Kim. 2022. Federated Semi-Supervised Learning with Prototypical Networks. *arXiv preprint arXiv:2205.13921* (2022).

[21] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[22] AJ Lawrance and PAW Lewis. 1977. An exponential moving-average sequence and point process (EMA1). *J. Appl. Probab.* 14, 1 (1977), 98–113.

[23] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, Vol. 3. 896.

[24] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. 2019. Federated learning for keyword spotting. In *ICASSP*. 6341–6345.

[25] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. 2022. PseCo: Pseudo Labeling and Consistency Training for Semi-Supervised Object Detection. *arXiv preprint arXiv:2203.16317* (2022).

[26] Ming Li, Qingli Li, and Yan Wang. 2023. Class Balanced Adaptive Pseudo Labeling for Federated Semi-Supervised Learning. In *CVPR*. 16292–16301.

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

[28] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *MLSys* 2 (2020), 429–450.

[29] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623* (2021).

[30] Xin-Chun Li, De-Chuan Zhan, Yunfeng Shao, Bingshuai Li, and Shaoming Song. 2021. Fedphp: Federated personalization with inherited private models. In *ECML/PKDD*. 587–602.

[31] Xiaoxiao Liang, Yiqun Lin, Huazhu Fu, Lei Zhu, and Xiaomeng Li. 2022. RSCFed: Random Sampling Consensus Federated Semi-supervised Learning. In *CVPR*. 10154–10163.

[32] Haowen Lin, Jian Lou, Li Xiong, and Cyrus Shahabi. 2021. Semifed: Semi-supervised federated learning with consistency and pseudo-labeling. *arXiv preprint arXiv:2108.09412* (2021).

[33] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020. Ensemble distillation for robust model fusion in federated learning. *NIPS* 33 (2020), 2351–2363.

[34] Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng. 2021. Federated semi-supervised medical image classification via inter-client relation matching. In *MICCAI*. 325–335.

[35] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. 2020. Unbiased Teacher for Semi-Supervised Object Detection. In *ICLR*.

[36] Zewei Long, Jiaqi Wang, Yaqing Wang, Houping Xiao, and Fenglong Ma. 2021. FedCon: A Contrastive Framework for Federated Semi-Supervised Learning. *arXiv preprint arXiv:2109.04533* (2021).

[37] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. 2021. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *NIPS* 34 (2021), 5972–5984.

[38] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*. 1273–1282.

[39] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).

[40] Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. 2022. How to measure uncertainty in uncertainty sampling for active learning. *Mach. Learn.* 111, 1 (2022), 89–122.

[41] Daniel Peterson, Pallika Kanani, and Virendra J Marathe. 2019. Private federated learning with domain adaptation. *arXiv preprint arXiv:1912.06733* (2019).

[42] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. 2021. Meta pseudo labels. In *CVPR*. 11557–11568.

[43] Jacob Poushter et al. 2016. Smartphone ownership and internet usage continues to climb in emerging economies. *Pew research center* 22, 1 (2016), 1–44.

[44] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. 2021. Data augmentation can improve robustness. *NIPS* 34 (2021), 29935–29948.

[45] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. 2021. Adaptive Federated Optimization. In *ICLR*.

[46] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *NIPS* 29 (2016).

[47] Tao Shen, Jie Zhang, Xinkang Jia, Fengda Zhang, Gang Huang, Pan Zhou, Kun Kuang, Fei Wu, and Chao Wu. 2020. Federated mutual learning. *arXiv preprint arXiv:2006.16765* (2020).

[48] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. 2018. Transductive semi-supervised deep learning using min-max features. In *ECCV*. 299–315.

[49] Yanhang Shi, Siguang Chen, and Haijun Zhang. 2022. Uncertainty Minimization for Personalized Federated Semi-Supervised Learning. *arXiv preprint arXiv:2205.02438* (2022).

[50] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NIPS* 33 (2020), 596–608.

[51] Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. 2021. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *CVPR*. 9311–9319.

[52] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NIPS* 30 (2017).

[53] Yu Wang, Pengchong Qiao, Chang Liu, Guoli Song, Xiawu Zheng, and Jie Chen. 2023. Out-of-Distributed Semantic Pruning for Robust Semi-Supervised Learning. In *CVPR*. 23849–23858.

[54] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. 2019. Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning. In *INFOCOM*. 2512–2520.

[55] Zhiguo Wang, Xintong Wang, Ruoyu Sun, and Tsung-Hui Chang. 2021. Federated Semi-Supervised Learning with Class Distribution Mismatch. *arXiv preprint arXiv:2111.00010* (2021).

[56] Wenqi Wei, Ling Liu, Margaret Loper, Ka-Ho Chow, Mehmet Emre Gursoy, Stacey Truex, and Yanzhao Wu. 2020. A framework for evaluating gradient leakage attacks in federated learning. *arXiv preprint arXiv:2004.10397* (2020).

[57] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *NIPS* 33 (2020), 6256–6268.

[58] Dong Yang, Ziyue Xu, Wenqi Li, Andriy Myronenko, Holger R Roth, Stephanie Harmon, Sheng Xu, Baris Turkbey, Evrim Turkbey, Xiaosong Wang, et al. 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. *Med. Image Anal.* 70 (2021), 101992.

[59] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903* (2018).

[60] Yuzhe Yang and Zhi Xu. 2020. Rethinking the value of labels for improving class-imbalanced learning. *NIPS* 33 (2020), 19290–19301.

[61] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *NIPS* 34 (2021), 18408–18419.

[62] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *CVPR*. 4320–4328.

[63] Zhengming Zhang, Yaoqing Yang, Zhewei Yao, Yujun Yan, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. 2021. Improving semi-supervised federated learning by reducing the gradient diversity of models. In *IEEE BigData*. 1214–1225.

[64] Zhen Zhao, Luping Zhou, Yue Duan, Lei Wang, Lei Qi, and Yinghuan Shi. 2022. Dc-ssl: Addressing mismatched class distribution in semi-supervised learning. In *CVPR*. 9757–9765.