Combating Data Imbalances in Federated Semi-supervised Learning with Dual Regulators

Sikai Bai[1]*, Shuaicheng Li[2*†], Weiming Zhuang[3‡], Jie Zhang[4‡], Kunlin Yang[2], Jun Hou[2‡], Shuai Zhang[2], Shuai Yi[2], Junyu Gao[5]

[1]The Hong Kong University of Science and Technology
[2]SenseTime Research
[3]Sony AI
[4]The Hong Kong Polytechnic University
[5]Northwestern Polytechnical University
{whitesk1973, gjy3035}@gmail.com, {lishuaicheng, yangkunlin, houjun, yishuai, zhangshuai}@sensetime.com, weiming001@e.ntu.edu.sg, jie-comp.zhang@polyu.edu.hk

Abstract

Federated learning has become a popular method to learn from decentralized heterogeneous data. Federated semi-supervised learning (FSSL) emerges to train models from a small fraction of labeled data due to label scarcity on decentralized clients. Existing FSSL methods assume independent and identically distributed (IID) labeled data across clients and consistent class distribution between labeled and unlabeled data within a client. This work studies a more practical and challenging scenario of FSSL, where data distribution is different not only across clients but also within a client between labeled and unlabeled data. To address this challenge, we propose a novel FSSL framework with dual regulators, FedDure. FedDure lifts the previous assumption with a coarse-grained regulator (C-reg) and a fine-grained regulator (F-reg): C-reg regularizes the updating of the local model by tracking the learning effect on labeled data distribution; F-reg learns an adaptive weighting scheme tailored for unlabeled instances in each client. We further formulate the client model training as bi-level optimization that adaptively optimizes the model in the client with two regulators. Theoretically, we show the convergence guarantee of the dual regulators. Empirically, we demonstrate that FedDure is superior to the existing methods across a wide range of settings, notably by more than 11% on CIFAR-10 and CINIC-10 datasets.

Introduction

Federated learning (FL) is an emerging privacy-preserving machine learning technique (McMahan et al. 2017), where multiple clients collaboratively learn a model under the coordination of a central server without exchanging private data. It has empowered a wide range of applications, including healthcare (Kaissis et al. 2020; Li et al. 2019), consumer products (Hard et al. 2018; Niu et al. 2020), and etc.

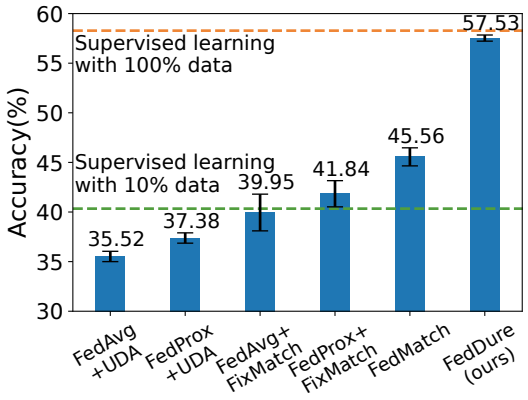


Figure 1: Existing federated semi-supervised learning (FSSL) methods cannot address heterogeneity between labeled and unlabeled data within a client (internal imbalance) and heterogeneous data across clients (external imbalance); some of them are even worse than supervised FL using 10% data (green line, which is FedAvg* in Table 1). Our proposed FedDure significantly outperforms existing methods. These experiments are based on three runs on CIFAR-10 and we provide more description in Section Experiments.

The majority of existing FL works (McMahan et al. 2017; Wang et al. 2020; Li, He, and Song 2021) assume that the private data in clients are fully labeled, but the assumption is unrealistic in real-world federated applications as annotating data is time-consuming, laborious, and expensive. To remedy these issues, federated semi-supervised learning (FSSL) is proposed to improve model performance with limited labeled and abundant unlabeled data on each client (Jin et al. 2020a). In particular, prior works (Jeong et al. 2021; Liu et al. 2021) have achieved competitive performance by exploring inter-client mutual knowledge. However, they usually focus on mitigating heterogeneous data distribution across clients (**external imbalance**) while as-

*Equal contribution.
†Project leader.
‡Corresponding author.

suming that labeled and unlabeled training data are drawn from the same independent and identical distribution. These assumptions enforce strict requirements of data annotation and would not be practical in many real-world applications. A general case is that labeled and unlabeled data are drawn from different distributions (**internal imbalance**). For example, photo gallery on mobile phones contains many more irrelevantly unlabeled images than the ones that are labeled manually for classification task (Yang et al. 2011).

Existing FSSL methods perform even worse than training with only a small portion of labeled data, under this realistic and challenging FSSL scenario with external and internal imbalances, as shown in Figure 1. The main reasons of performance degradation are two-fold: 1) internal imbalance leads to intra-client skewed data distribution, resulting in heterogeneous local training; 2) external imbalance leads to inter-client skewed data distribution, resulting in client drift (Charles and Konečnỳ 2021; Karimireddy et al. 2020). The co-occurrence of internal and external data imbalances amplifies the impact of client drifts and local inconsistency, leading to performance degradation.

To address the above issues, we propose a new federated semi-supervised learning framework termed **FedDure**. FedDure explores two adaptive regulators, a coarse-grained regulator (**C-reg**) and a fine-grained regulator (**F-reg**), to flexibly update the local model according to the learning process and outcome of the client's data distributions. Firstly, C-reg regularizes the updating of the local model by tracking the learning effect on labeled data. By utilizing the real-time feedback from C-reg, FedDure rectifies inaccurate model predictions and mitigates the adverse impact of internal imbalance. Secondly, F-reg learns an adaptive weighting scheme tailored for each client; it automatically equips a *soft* weight for each unlabeled instance to measure its contribution. This scheme automatically adjusts the instance-level weights to strengthen (or weaken) its confidence according to the feedback of F-reg on the labeled data to further address the internal imbalance. Besides, FedDure mitigates the client drifts caused by external imbalance by leveraging the global server model to provide guidance knowledge for C-reg. During the training process, FedDure utilizes the bi-level optimization strategy to alternately update the local model and dual regulators in local training. Figure 1 shows that FedDure significantly outperforms existing methods and its performance is even close to fully supervised learning (orange line) under internal and external imbalance. To the end, the main contributions are three-fold:

- We are the first work that investigates a more practical and challenging scenario of FSSL, where data distribution differs not only across clients (external imbalance) but between labeled and unlabeled data within a client (internal imbalance).
- We propose FedDure, a new FSSL framework that designs dual regulators to adaptively update the local model according to the unique learning processes and outcomes of each client.
- We theoretically analyze the convergence of dual regulators and empirically demonstrate that FedDure is superior to the state-of-the-art FSSL approaches across multiple benchmarks and data settings, improving accuracy by 12.17% on CIFAR10 and by 11.16% on CINIC-10 under internal and external imbalances.

Related Work

Federated Learning (FL) is an emerging distributed training technique that trains models on decentralized clients and aggregates model updates in a central server (Yang et al. 2019). FedAvg (McMahan et al. 2017) is a pioneering work that aggregates local models updated by weighted averaging. Statistical heterogeneity is an important challenge of FL in real-world scenarios, where the data distribution is inconsistent among clients (Li et al. 2020a), which can result in drift apart between global and local model, i.e., client-drift (Charles and Konečnỳ 2021). A plethora of works have been proposed to address this challenge with approaches like extra data sharing, regularization, new aggregation mechanisms, and personalization (Zhao et al. 2018; Li et al. 2022a, 2021b; Xu et al. 2021). These approaches commonly consider only supervised learning settings and may not be simply applied to scenarios where only a small portion of data is labeled. Moreover, some work studies un/self-supervised learning settings (Zhuang, Wen, and Zhang 2022; Wang et al. 2021; Li et al. 2021a) to learn generic representations with purely unlabeled data on clients, and these methods require IID labeled data for fine-tuning the representations for downstream tasks (Li et al. 2022b; Bai et al. 2021). Our work primarily focuses on federated semi-supervised learning, where a small fraction of data has labels in each client.

Semi-Supervised Learning aims to utilize unlabeled data for performance improvements and is usually divided into two popular branches pseudo labeling and consistency regularization. Pseudo-labeling methods (Lee et al. 2013; Zou et al. 2022; Pham et al. 2021; Chen et al. 2022) usually generate artificial labels of unlabeled data from the model trained by labeled data and apply the filtered high-confidence labels as supervised signals for unlabeled data training. MPL (Pham et al. 2021) extends the knowledge distillation to SSL by optimizing the teacher model with feedback from the student model. Consistency regularization approaches (Lee et al. 2022; Tarvainen and Valpola 2017) regularize the outputs of different perturbed versions of the same input to be consistent. Many works (Sohn et al. 2020; Zhang et al. 2021a; Lee et al. 2022) apply data augmentation as a perturbed strategy for pursuing outcome consistency.

Federated Semi-Supervised Learning (FSSL) considers learning models from decentralized clients where a small amount of labeled data resides on either clients or the server (Jin et al. 2020b). FSSL scenarios can be classified into three categories: (1) Labels-at-Server assumes that clients have purely unlabeled data and the server contains some labeled data (Lin et al. 2021; He et al. 2021; Zhang et al. 2021b; Diao, Ding, and Tarokh 2021); (2) Labels-at-Clients considers each client has mostly unlabeled data and a small amount of labeled data (Jeong et al. 2021); (3) Labels-at-Partial-Clients assumes that the majority of clients contain fully unlabeled data while numerous clients have fully labeled data

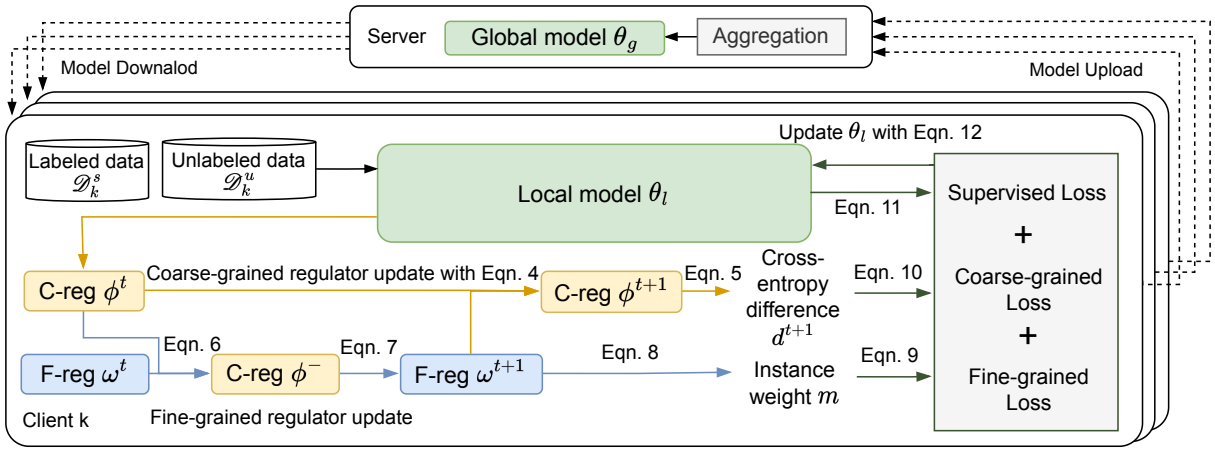


Figure 2: Illustration of Federated Semi-Supervised Learning Framework with Dual Regulator (FedDure). FedDure contains a coarse-grained regulator (C-reg) and a fine-grained regulator (F-reg) to adaptively guide local updates in each client: C-reg dynamically regulates the importance of local training on the unlabeled data by reflecting the overall learning effect on labeled data; F-reg regulates the performance contribution of each unlabeled sample.

(Lin et al. 2021; Liang et al. 2022). Labels-at-Clients has been largely overlooked; prior work (Jeong et al. 2021) proposes inter-client consistency loss, but it shares extra information among clients and bypasses the internal class imbalance issue. This work introduces dual regulators to address the issue, without extra information shared among clients.

Method

This section first defines the problem and introduces a novel framework with dual regulators (FedDure). Using dual regulators, we then build a bi-level optimization strategy for federated semi-supervised learning.

Problem Definition

We focus on Federated Semi-Supervised Learning (FSSL) with external and internal imbalance problems. Specifically, we assume that there are K clients, denoted as $\{\mathcal{C}_1, ..., \mathcal{C}_K\}$. Federated learning aims to train a generalized global model f_g with parameter θ_g. It coordinates decentralized clients to train their local models $\mathcal{F}_l = \{f_{l,1}, ..., f_{l,K}\}$ with parameters $\{\theta_{l,1}, ..., \theta_{l,K}\}$, where each client is only allowed to access its own local private dataset. In the standard semi-supervised setting, the dataset contains a labeled set $\mathcal{D}^s = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N^s}$ and an unlabeled set $\mathcal{D}^u = \{\mathbf{u}_i\}_{i=1}^{N^u}$, where $N^s \ll N^u$. Under FSSL, the private dataset $\mathcal{D}_k$ of each client C_k contains N_k^s labeled instances $\mathcal{D}_k^s = \{\mathbf{x}_{i,k}, \mathbf{y}_{i,k}\}_{i=1}^{N_k^s}$ and N_k^u unlabeled instances $\mathcal{D}_k^u = \{\mathbf{u}_{i,k}\}_{i=1}^{N_k^u}$. The internal imbalance means that the distribution of $\mathcal{D}_k^s$ and $\mathcal{D}_k^u$ are different; the external imbalance refers to different distributions between D_k in different clients k. We provide a detailed description in Subsection Data Heterogeneity.

In this work, we primarily focus on image datasets. For an unlabeled image $\mathbf{u}_k$ in client C_k, we compute the corresponding pseudo label $\hat{\mathbf{y}}_k$ with the following equation:

$$\hat{\mathbf{y}}_k = \text{argmax}(f_{l,k}(\mathcal{T}_w(\mathbf{u}_k); \boldsymbol{\theta}_{l,k})), \tag{1}$$

where $\mathcal{T}_w(\mathbf{u}_k)$ is the weakly-augmented version of $\mathbf{u}_k$ and the pseudo labeling dataset in the client C_k is denoted as $\mathcal{D}_k^u = \{\mathbf{u}_{i,k}, \hat{\mathbf{y}}_{i,k}\}_{i=1}^{N_k^u}$. We omit the client index k in the parameters later for simplicity of notation.

Dual Regulators

In this section, we present federated semi-supervised learning with dual regulator, termed FedDure. It dynamically adjusts gradient updates in each client according to the class distribution characteristics with two regulators, a coarse-grained regulator (C-reg) and a fine-grained regulator (F-reg). Figure 2 depicts the optimization process with these two regulators. We introduce the regulators and present the optimization process in the following subsections.

Coarse-grained Regulator (C-reg). Existing FSSL methods decompose the optimization on the labeled and unlabeled data, leading to heterogeneous local training. C-reg remedies the challenge with a collaborative training manner. Intuitively, the parameters of the local model can be rectified according to the feedback from C-reg, which dynamically regulates the importance of local training on all unlabeled data by quantifying the overall learning effect using labeled data. It contributes to counteracting the adverse impact introduced by internal imbalance and preventing corrupted pseudo-labels (Chen et al. 2022). Meanwhile, C-reg acquires global knowledge by initializing with the received server model parameters at the beginning of each round of local training, which can provide global guidance to the local model to mitigate external imbalance (client-drift).

We define C-reg as f_d with parameters $\boldsymbol{\phi}$. At training iteration t, C-reg searches its optimal parameter $\boldsymbol{\phi}^*$ by minimizing the cross-entropy loss on unlabeled data with pseudo labels. Actually, the optimal parameter $\boldsymbol{\phi}^*$ is related to the

local model's parameter $\boldsymbol{\theta}_l$ via the generated pseudo label, and we denote the relationship as $\boldsymbol{\phi}^*(\boldsymbol{\theta}_l)$. Since it requires heavy computational costs to explore the optimal parameter $\boldsymbol{\phi}^*$, we approximate $\boldsymbol{\phi}^*$ by performing *one gradient step* $\boldsymbol{\phi}^{t+1}$ at training iteration t (i.e., $\boldsymbol{\phi}^t$).

Practically, we introduce the updated fine-grained regulator (F-reg) to measure the scalar weight for each unlabeled instance for updating C-reg. The formulation to optimize C-reg is as follows:

$$\boldsymbol{\phi}^{t+1} = \boldsymbol{\phi}^t - \eta_s \nabla_{\boldsymbol{\phi}^t} \mathbb{E}_{\boldsymbol{u}} \mathcal{H}(\boldsymbol{w}^{t+1}; \boldsymbol{\phi}^t) \mathcal{L}_{ce}\left(\hat{\mathbf{y}}, f_d\left(\mathcal{T}_s(\mathbf{u}); \boldsymbol{\phi}^t\right)\right), \tag{2}$$

where $\mathcal{H}(\boldsymbol{w}^{t+1}; \boldsymbol{\phi}^t) = f_w\left(f_d\left(\mathcal{T}_s(\mathbf{u}); \boldsymbol{\phi}^t\right); \boldsymbol{w}^{t+1}\right)$, f_w is the fine-grained regulator (F-reg), and $\boldsymbol{w}^{t+1}$ is the parameters of F-reg updated by Eqn. 5, which is detailed in the following subsection. $\mathcal{T}_s(\mathbf{u})$ is the strongly-augmented unlabeled image $\mathbf{u}$ and $f_d(\mathcal{T}_s(\mathbf{u}); \boldsymbol{\phi}^t)$ is the output vector of f_d to evaluate the quality of pseudo labels from the local model.

Next, we quantify the learning effect of the local model with the C-reg using labeled samples by computing the cross-entropy difference d^{t+1} of C-reg between training iterations t and $t+1$:

$$d^{t+1} = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\mathcal{L}_{ce}(\mathbf{y}, f_d(\mathbf{x}; \boldsymbol{\phi}^t)) - \mathcal{L}_{ce}(\mathbf{y}, f_d(\mathbf{x}; \boldsymbol{\phi}^{t+1}))\right]. \tag{3}$$

The quantized learning effect is further used as the reward information to optimize the local model by regulating the importance of local training on unlabeled data. In particular, the cross-entropy differences d^{t+1} signify the generalization gap for the C-reg updated by the pseudo labels from the local model.

Fine-grained Regulator (F-reg). Previous SSL methods usually utilize a *fixed threshold* to filter noisy pseudo labels (Sohn et al. 2020), but they are substantially hindered by corrupted labels or class imbalance on unlabeled data. Internal and external imbalances in FSSL could amplify these problems, leading to performance degradation. To tackle the challenge, F-reg regulates the importance of each unlabeled instance in local training for mitigating the learning bias caused by internal imbalance. It learns an adaptive weighting scheme tailored for each client according to unlabeled data distribution. A unique weight is generated for each unlabeled image to measure the contribution of the image to overall performance. We construct F-reg f_w parameterized by $\boldsymbol{w}$[1]. Before updating F-reg, we perform *one gradient step* update of C-reg $\boldsymbol{\phi}$ to associate F-reg and C-reg:

$$\boldsymbol{\phi}^- = \boldsymbol{\phi}^t - \eta_s \nabla_{\boldsymbol{\phi}^t} \mathbb{E}_{\boldsymbol{u}} \mathcal{H}(\boldsymbol{w}^t; \boldsymbol{\phi}^t) \mathcal{L}_{ce}\left(\hat{\mathbf{y}}, f_d\left(\mathcal{T}_s(\mathbf{u}); \boldsymbol{\phi}^t\right)\right), \tag{4}$$

where *one gradient step* of C-reg $\boldsymbol{\phi}^-$ depends on the F-reg $\boldsymbol{w}^t$ and regards the others as fixed parameters. Next, we optimize F-reg in local training iteration t, where the optimal parameter $\boldsymbol{w}^*$ is approximated by one gradient step of F-reg (i.e., $\boldsymbol{w}^{t+1}$). The optimization of F-reg is formulated as:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta_w \nabla_{\boldsymbol{w}^t} \mathbb{E}_{\mathbf{x},\mathbf{y}} \mathcal{L}_{ce}\left(\mathbf{y}, f_d(\mathbf{x}; \boldsymbol{\phi}^-(\boldsymbol{w}^t))\right), \tag{5}$$

[1]F-reg is a MLP architecture with one fully connected layer with 128 filters and a Sigmoid function.

where $f_d(\mathbf{x}; \boldsymbol{\phi}^-(\boldsymbol{w}^t))$ is the output of f_d on labeled data. We then introduce a re-weighting scheme that calculates a unique weight $\boldsymbol{m}_i$ for i-th unlabeled sample:

$$\boldsymbol{m}_i = f_w(f_l(\mathcal{T}_s(\mathbf{u}_i), \boldsymbol{\theta}_l^t), \boldsymbol{w}^{t+1}). \tag{6}$$

Note that $\boldsymbol{m}_i$ is a scalar to re-weight the importance of the corresponding unlabeled image.

Bi-level Optimization

In this section, we present optimization processes for the dual regulators and local model θ. We alternatively train two regulators, which approximate a gradient-based bi-level optimization procedure (Finn, Abbeel, and Levine 2017; Liu, Simonyan, and Yang 2018). Then, we update the local model with fixed C-reg and F-reg.

Update F-reg. Firstly, we obtain one gradient step update of C-reg $\boldsymbol{\phi}^-$ using Eqn. 4. After that, the supervised loss $\mathcal{L}_{ce}\left(\mathbf{y}, f_d(\mathbf{x}; \boldsymbol{\phi}^-(\boldsymbol{w}^t))\right)$ guides the update of the F-reg with Eqn. 5. Since $\boldsymbol{w}^t$ is explicitly beyond the supervised loss, the updating of F-reg can be achieved by standard backpropagation using the chain rule.

Update C-reg. After updating the parameters of F-reg, we update C-reg by Eqn. 2, regarding local model $\boldsymbol{\theta_l}^t$ as fixed parameters.

Update Local Model with F-reg. We use the updated F-reg $\boldsymbol{w}^{t+1}$ to calculate a unique weight $\boldsymbol{m}_i$ for i-th unlabeled sample with Eqn. 6. The gradient optimization is formulated as:

$$\boldsymbol{g}_u^t = \mathbb{E}_{\mathbf{u}}\left[\nabla_{\boldsymbol{\theta}_l^t} \mathcal{L}_{ce}\left(\hat{\mathbf{y}}, f_l\left(\mathcal{T}_s(\mathbf{u}); \boldsymbol{\theta}_l^t\right)\right) \cdot \boldsymbol{m}\right]. \tag{7}$$

Update Local Model with C-reg. We then use C-reg to calculate entropy difference d^{t+1} in Eqn. 3. The entropy difference d^{t+1} is adopted as a reward coefficient to adjust the gradient update of the local model on unlabeled data. The formulation is as follows:

$$\boldsymbol{g}_d^t = d^{t+1} \cdot \nabla_{\boldsymbol{\theta}_l^t} \mathbb{E}_{\mathbf{u}} \mathcal{L}_{ce}\left(\hat{\mathbf{y}}, f_l\left(\mathcal{T}_s(\mathbf{u}); \boldsymbol{\theta}_l^t\right)\right), \tag{8}$$

where this learning process is derived from a meta-learning strategy, provided in supplementary materials about the proof for analysis.

Update Local Model with Supervised Loss. Besides, we compute the gradient local model on labeled data as:

$$\boldsymbol{g}_s^t = \nabla_{\boldsymbol{\theta}_l^t} \mathbb{E}_{\mathbf{x},\mathbf{y}} \mathcal{L}_{ce}\left(\mathbf{y}, f_l\left(\mathbf{x}; \boldsymbol{\theta}_l^t\right)\right). \tag{9}$$

On this basis, we update the local model's parameter with the above gradient computation in Eqn. 7, 8 and 9, which is defined as:

$$\boldsymbol{\theta}_l^{t+1} = \boldsymbol{\theta}_l^t - \eta\left(\boldsymbol{g}_s^t + \boldsymbol{g}_u^t + \boldsymbol{g}_d^t\right), \tag{10}$$

where η denotes the learning rate of the local model. Finally, after T local epochs, the local model is returned to the central server. The server updates global model $\boldsymbol{\theta_g}^{r+1}$ by weighted averaging the parameters from these received local models in the current round, and $r+1$-th round is conducted by sending $\boldsymbol{\theta_g}^{r+1}$ to the randomly selected clients as initialization. We present the pipeline of the overall optimization process in the supplementary material about the proof for analysis.

Convergence of Optimization Process

In this section, we further analyze the convergence of our optimizations. When updating F-reg in Eqn. 5, $\boldsymbol{w}^t$ is explicitly beyond the supervised loss, the optimization of F-reg can be easily implemented by automatic backpropagation using the chain rule. We only discuss the convergence of the bi-level optimizations using the meta-learning process.

Update Local Model with C-reg. The local model tries to update its parameters on the feedback from the updated coarse-grained regulator (C-reg), which adjusts the learning effect via the meta-learning process. The cross-entropy loss on labeled data $\mathcal{L}_{ce}(\mathbf{y}, f_d(\mathbf{x}; \boldsymbol{\phi}^{t+1}(\boldsymbol{\theta}_l^t))$ is applied to characterize the quality of learning effect from the local model. The CE loss function is related to $\boldsymbol{\theta}_l^t$.

Theorem 1 *Suppose that supervised loss function $\mathcal{L}_{ce}(\boldsymbol{y}, f_d(\boldsymbol{x}; \boldsymbol{\phi}^{t+1}(\boldsymbol{\theta}_l^t))$ is L-Lipschitz and has ρ-bounded gradients. The $\mathcal{L}_{ce}\left(\hat{\boldsymbol{y}}, f_d\left(\mathcal{T}_s(\boldsymbol{u}); \boldsymbol{\phi}^t\right)\right)$ has ρ-bounded gradients and twice differential with Hessian bounded by $\mathcal{B}$. Let the learning rate $\eta_s = \min\{1, \frac{e}{T}\}$ for constant $e > 0$, and $\eta = \min\{\frac{1}{L}, \frac{c}{\sqrt{T}}\}$ for some $c > 0$, such that $\frac{\sqrt{T}}{c} \geq L$. Thus, the optimization of the local model using coarse-grained regulator can achieve:*

$$\min_{0 \leq t \leq T} \mathbb{E}[\|\nabla_{\theta_l} \mathcal{L}_{ce}(\boldsymbol{y}, f_d(\boldsymbol{x}; \boldsymbol{\phi}^{t+1}(\boldsymbol{\theta}_l^t))\|_2^2] \leq \mathcal{O}(\frac{c}{\sqrt{T}}). \quad (11)$$

Update C-reg. We introduce updated F-reg to measure the contributions of each instance for updating C-reg in Eqn. 2, where $\mathcal{H}(\boldsymbol{w}^{t+1}; \boldsymbol{\phi}^t)$ is related to $\boldsymbol{\phi}^t$. The updated F-reg adjusts the learning contributions on each unlabeled instance for regulating the optimization of C-reg. We conclude that our C-reg can always achieve convergence when introducing the feedback from F-reg.

Theorem 2 *Suppose supervised and unsupervised loss functions are Lipschitz-smooth with constant L and have ρ-bounded gradient. The $\mathcal{H}(\cdot)$ is differential with a ϵ-bounded gradient and twice differential with its Hessian bounded by $\mathcal{B}$. Let learning rate η_s satisfies $\eta_s = \min\{1, \frac{k}{T}\}$ for constant $k > 0$, such that $\frac{k}{T} < 1$. $\eta_w = \min\{\frac{1}{L}, \frac{c}{\sqrt{T}}\}$ for constant $c > 0$ such that $\frac{\sqrt{T}}{c} \geq L$. The optimization of the coarse-grained regulator can achieve:*

$$\lim_{t \to \infty} \mathbb{E}[\|\nabla_\phi \mathcal{H}(\boldsymbol{w}^{t+1}; \boldsymbol{\phi}^t) \mathcal{L}_{ce}\left(\hat{\boldsymbol{y}}, f_d\left(\mathcal{T}_s(\boldsymbol{u}); \boldsymbol{\phi}^t\right)\right)\|_2^2] = 0. \quad (12)$$

Experiments

In this section, we demonstrate the effectiveness and robustness of our method through comprehensive experiments in three benchmark datasets under multiple data settings.

Experimental Setup

Datasets. We conduct comprehensive experiments on three datasets, including CIFAR-10 (Krizhevsky, Hinton et al. 2009), Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017) and CINIC-10 (Darlow et al. 2018). All datasets are split according to official guidelines; we provide more dataset descriptions and split strategies in the supplementary material.

Data Heterogeneity. We construct three data heterogeneity settings with different data distributions. We denote each setting as ($\mathcal{A}$, $\mathcal{B}$), where $\mathcal{A}$ and $\mathcal{B}$ are data distribution of labeled and unlabeled data, respectively. The settings are as follows: (1) ***(IID, IID)*** means both labeled and unlabeled data are IID. By default, we use 5 instances per class to build the labeled dataset for each client. The remaining instances of each class are divided into K clients evenly to build an unlabeled dataset. (2) ***(IID, DIR)*** means labeled data is the same as (IID, IID), but the unlabeled data is constructed with Dirichlet distribution to simulate data heterogeneity, where each client could only contain a subset of classes. (3) ***(DIR, DIR)*** constructs both labeled and unlabeled data with Dirichlet distribution. It simulates external and internal class imbalance, where the class distributions across clients and within a client are different. We allocate 500 labeled data per class to 100 clients using the Dirichlet process. The rest instances are divided into each client with another Dirichlet distribution. Figure 3 compares the data distribution of FedMatch (Batch NonIID) (Jeong et al. 2021) and ours. Our (DIR, DIR) setting presents class imbalance across clients (external imbalance) and between labeled and unlabeled data within a client (internal imbalance).

Implementation Details. We use the Adam optimizer with momentum $= 0.9$, batch size $= 10$ and learning rates $= 0.0005$ for η_s, η and η_w. If there is no specified description, our default settings also include local iterations $T = 1$, the selected clients in each round $S = 5$, and the number of clients $K = 100$. For the DIR data configuration, we use a Dirichlet distribution $Dir(\gamma)$ to generate the DIR data for all clients, where $\gamma = 0.5$ for all three datasets. We adopt the ResNet-9 network as the default backbone architecture for local models and the coarse-grained regulator, while an MLP is utilized for the fine-grained regulator.

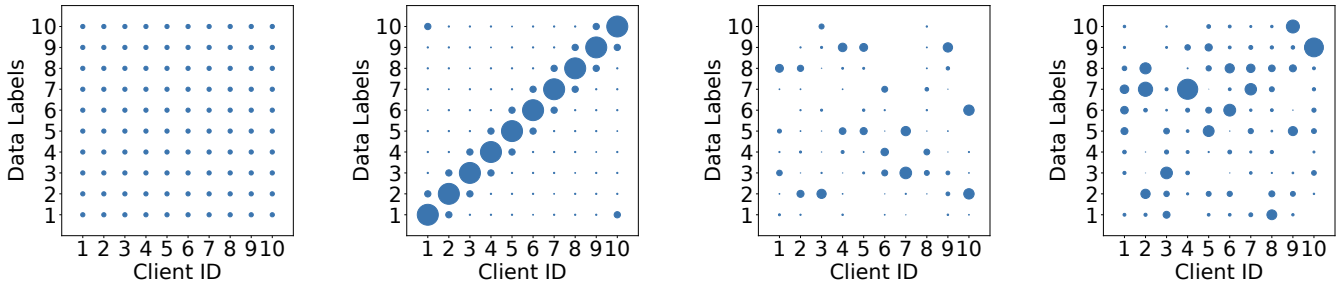
Baselines. We compare the following methods in experiments. ***FedAvg**** denotes FedAvg (McMahan et al. 2017) only trained on labeled samples in FSSL (about 10% data). ***FedAvg-SL*** and ***FedProx-SL*** are fully supervised training using FedAvg (McMahan et al. 2017) and FedProx (Li et al. 2020b), respectively. ***FedAvg+UDA***, ***FedProx+UDA***, ***FedAvg+Fixmatch***, and ***FedProx+Fixmatch***: a naive combination between semi-supervised methods (UDA (Xie et al. 2020) and Fixmatch (Sohn et al. 2020)) and FL algorithms. They use labeled and unlabeled data, but need to specify a predefined threshold on pseudo labels. ***FedMatch*** (Jeong et al. 2021) adopts inter-consistency loss, and disjoint loss for model training which is the state-of-the-art FSSL method. Note that we use the same hyper-parameters for FedDure and other methods in all experiments.

Performance Comparison

Table 1 reports the overall results of FedDure and other state-of-the-art methods on the three datasets. These results are averaged over 3 independent runs. Our FedDure achieves state-of-the-art FSSL performances on all datasets and data settings. ***(IID, IID) setting:*** compared with the naive combination of FSSL methods and FedMatch, our FedDure significantly outperforms them on all three datasets. Specifically, when evaluated on CINIC-10, which is a more diffi-

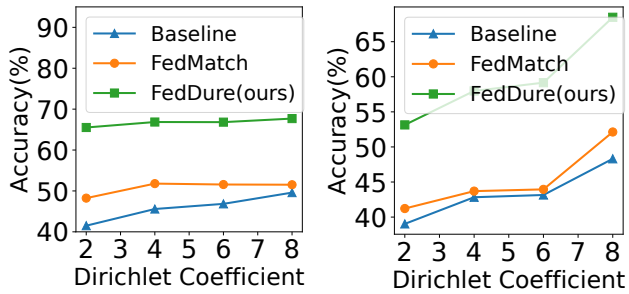
Methods	CIFAR10			Fashion-MNIST			CINIC-10		
	(IID, IID)	(IID, DIR)	(DIR, DIR)	(IID, IID)	(IID, DIR)	(DIR, DIR)	(IID, IID)	(IID, DIR)	(DIR, DIR)
FedAvg*	45.68	43.83	40.34	85.56	84.84	82.24	40.73	39.00	28.09
FedAvg-SL	75.47	66.70	58.38	89.87	88.60	86.95	67.97	57.72	46.21
FedProx-SL	74.67	66.78	59.55	89.53	88.35	87.32	68.13	58.67	52.09
FedAvg+UDA	47.47	43.89	35.52	86.20	85.35	81.07	42.25	39.93	29.27
FedProx+UDA	46.49	42.82	37.38	84.78	84.50	82.94	41.81	39.40	33.26
FedAvg+Fixmatch	46.71	45.58	39.95	86.46	85.42	81.07	40.40	39.66	31.99
FedProx+Fixmatch	47.60	43,39	41.85	86.31	85.18	83.68	41.46	40.02	32.21
FedMatch	51.52	51.59	45.56	85.71	85.55	85.13	43.73	41.82	35.27
FedDure (Ours)	**67.69**	**66.85**	**57.73**	**88.69**	**88.21**	**86.96**	**56.36**	**55.10**	**46.43**

Table 1: Performance comparison of our proposed FedDure with state-of-the-art methods on three different data heterogeneity settings. FedDure outperforms all other methods in all settings.



(a) FedMatch: Labeled (b) FedMatch: Unlabeled (c) FedDure: Labeled (d) FedDure: Unlabeled

Figure 3: Comparison of data distribution between FedMatch (Jeong et al. 2021) and our (DIR, DIR) setting: (a) and (b) are labeled and unlabeled data distribution used in FedMatch, respectively; our data distribution in (c) and (d) present external imbalance across clients and internal imbalance between labeled and unlabeled data inside a client.



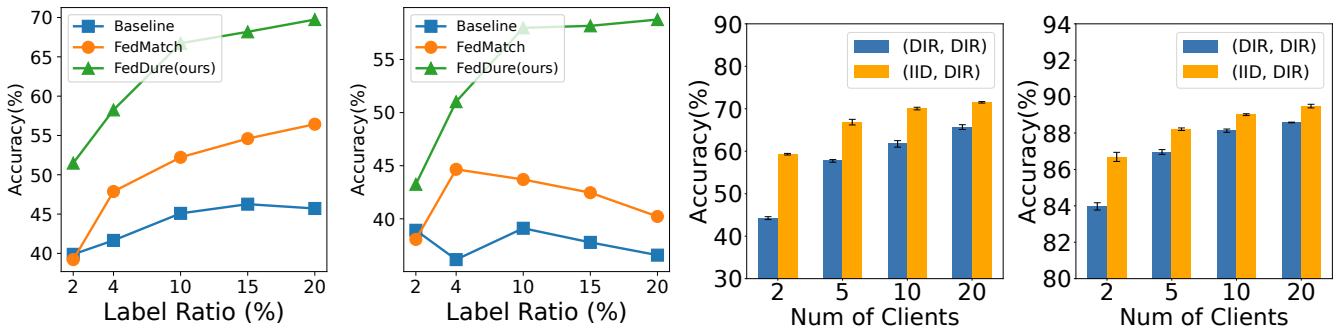
(a) Setting: (IID, DIR) (b) Setting: (DIR, DIR)

Figure 4: Impact of different Dirichlet coefficients under (IID, DIR) and (DIR, DIR) settings on CIFAR10 dataset.

cult dataset with a larger amount of unlabeled samples, other methods suffer from the performance bottleneck and are inferior on CIFAR-10 with fewer unlabeled samples. These results show that FedDure effectively alleviates the negative influence of mass unlabeled data by regulating the local model's optimization on unlabeled data through knowledge feedback from labeled data using F-reg and C-reg. ***(IID, DIR) setting:*** our FedDure is slightly affected by weak class mismatch on unlabeled data, but it significantly outperforms by FedMatch 15.26% on CIFAR10 dataset. Also, competitive performance is achieved compared to the supervised method FedAvg-SL on Fashion-MNIST. ***(DIR, DIR) setting:*** Under this more challenging and realistic setting, our FedDure significantly outperforms others by at least 11% on CIFAR-10 and CINIC-10 datasets. In particular, the performance of other methods drops dramatically, and in CIFAR10 and Fashion-MNIST datasets, some semi-supervised algorithms are even worse than FedAvg*. It means that unlabeled data might hurt performance due to the distribution mismatch between labeled and unlabeled data.

Ablation Study

Effectiveness of Components. To measure the importance of proposed components in our FedDure, we conduct ablation studies with the following variants in Table 2. (1) Baseline: the naive combination of FedAvg (McMahan et al. 2017) and Fixmatch (Sohn et al. 2020). (2) Baseline+MAML: an adaptive optimization for baseline based on vanilla meta-learning. The performance improvement over the baseline verifies the insights and the effectiveness of adopting client-specific optimization strategies via meta-learning. (3) Ours w F-reg: this variant denotes our FedDure removes the C-reg (i.e. g_d in Eqn.10) and updates F-reg with



(a) Setting: (IID, DIR) (b) Setting: (DIR, DIR) (c) CIFAR-10 (d) Fashion-MNIST

Figure 5: Analysis of the impacts of the number of labeled data and selected clients. (a) and (b) illustrate that FedDure consistently outperforms FedMatch and Baseline (FedAvg-Fixmatch) using different percentages of labeled data. (c) and (d) show that FedDure scales with increasing numbers of selected clients on CIFAR-10 and Fashion-MNIST datasets.

Ablated components	CIFAR-10		
	(DIR, DIR)	(IID, DIR)	(IID, IID)
Baseline	39.95	46.67	47.60
Baseline+MAML	47.69	56.05	59.78
F-reg	54.79	64.98	65.41
Avg F-reg	50.08	58.26	59.13
C-reg	56.46	65.92	66.54
Avg C-reg	52.32	60.07	62.07
FedDure	**57.73**	**66.85**	**67.69**

Table 2: Quantitative analysis of components of FedDure on CIFAR-10 and Fashion-MNIST datasets.

the local model. (4) Ours *w* C-reg: this variant indicates our FedDure replaces the dynamic weight (i.e. g_u in Eqn.10) and uses a fixed threshold to filter low-confidence pseudo labels. The performance advantage over Baseline+MAML shows that the two components are both more effective for FSSL scenarios. Moreover, C-reg can further make a performance boost under almost all data sets on CIFAR-10. This is because C-reg has targeted overall knowledge in local training, but there is no significant difference between the two components. (5) Avg C-reg and Avg F-reg: we set F-reg and C-reg to the average of corresponding regulators in previous rounds. Compared with F-reg and C-reg, the decrease in performance suggests the importance of online client-specific adaptive optimization in dual regulators.

Impacts of Data Heterogeneity. To demonstrate the robustness of our method against data imbalance, we characterize different levels of imbalances by Dirichlet distribution with different coefficients $\{0.3, 0.5, 0.7, 1.0\}$ and evaluate multiple methods. As illustrated in Figure 4(a) and 4(b), our FedDure consistently showcases substantial performance improvements across different levels of data imbalances. However, FedMatch and baseline (FedAvg-Fixmatch) suffer from rapid performance degradation when confronted with the higher data heterogeneous *(small coefficient)*

Number of Label Data per Client. We evaluate FedDure under the different percentages of labeled instances in each client in $\{2\%, 4\%, 10\%, 15\%, 20\%\}$. As illustrated in Figure 5(a) and 5(b), FedDure gains steady performance improvements with the number of labeled data increases in two data settings. In contrast, the baseline's performance remains relatively stagnant across both scenarios. As for FedMatch, a noticeable decline becomes obvious when the labeling ratio exceeds 4% in the (DIR, DIR) setting. These insightful findings underscore the efficacy of our dual regulators.

Number of Selected Clients per Round. Lastly, we investigate the performance on the impact of the number of selected clients varied in $\{2, 5, 10, 20\}$. As shown in Figure 5(c) and 5(d), significant improvements can be achieved by increasing the selected clients. Nevertheless, performance gains plateau as the chosen clients surpass a certain threshold. Our contention is that while the number of selected clients does exhibit a positive correlation with overall performance, our method delves into the intrinsic knowledge of each client to enhance the central server's overall performance. In scenarios where there are ample clients, our approach assimilates comprehensive knowledge, resulting in saturated performance.

Conclusion

In this paper, we introduce a more practical and challenging scenario of FSSL, data distribution is different across clients (external imbalance) and within a client (internal imbalance). We then design a new federated semi-supervised learning framework with dual regulators, FedDure, to address the challenge. Particularly, we propose a coarse-grained regulator (C-reg) to regularize the gradient update in client model training and present a fine-grained regulator (F-reg) to learn an adaptive weighting scheme for unlabeled instances for gradient update. Furthermore, we formulate the learning process in each client as bi-level optimization that optimizes the local model in the client adaptively and dynamically with these two regulators. Theoretically, we show the convergence guarantee of the regulators. Empirically, extensive experiments demonstrate the significance and effectiveness of FedDure.

Acknowledgments

This research was supported by fundings from the Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400003), Hong Kong RGC Research Impact Fund (No. R5060-19, No. R5034-18), Areas of Excellence Scheme (AoE/E-601/22-R), General Research Fund (No. 152203/20E, 152244/21E, 152169/22E, 152228/23E), Shenzhen Science and Technology Innovation Commission (JCYJ20200109142008673).

References

Bai, S.; Gao, J.; Wang, Q.; and Li, X. 2021. Multi-Domain Synchronous Refinement Network for Unsupervised Cross-Domain Person Re-Identification. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Charles, Z.; and Konečnỳ, J. 2021. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, 2575–2583. PMLR.

Chen, B.; Jiang, J.; Wang, X.; Wan, P.; Wang, J.; and Long, M. 2022. Debiased Self-Training for Semi-Supervised Learning. In *Advances in Neural Information Processing Systems*.

Darlow, L. N.; Crowley, E. J.; Antoniou, A.; and Storkey, A. J. 2018. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*.

Diao, E.; Ding, J.; and Tarokh, V. 2021. SemiFL: Communication efficient semi-supervised federated learning with unlabeled clients. *arXiv preprint arXiv:2106.01432*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Hard, A.; Rao, K.; Mathews, R.; Ramaswamy, S.; Beaufays, F.; Augenstein, S.; Eichner, H.; Kiddon, C.; and Ramage, D. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

He, C.; Yang, Z.; Mushtaq, E.; Lee, S.; Soltanolkotabi, M.; and Avestimehr, S. 2021. Ssfl: Tackling label deficiency in federated learning via personalized self-supervision. *arXiv preprint arXiv:2110.02470*.

Jeong, W.; Yoon, J.; Yang, E.; and Hwang, S. J. 2021. Federated semi-supervised learning with inter-client consistency & disjoint learning.

Jin, Y.; Wei, X.; Liu, Y.; and Yang, Q. 2020a. A survey towards federated semi-supervised learning. *arXiv preprint arXiv:2002.11545*, 50.

Jin, Y.; Wei, X.; Liu, Y.; and Yang, Q. 2020b. Towards utilizing unlabeled data in federated learning: A survey and prospective. *arXiv preprint arXiv:2002.11545*.

Kaissis, G. A.; Makowski, M. R.; Rückert, D.; and Braren, R. F. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lee, D.; Kim, S.; Kim, I.; Cheon, Y.; Cho, M.; and Han, W.-S. 2022. Contrastive Regularization for Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3911–3920.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10713–10722.

Li, S.; Cao, Q.; Liu, L.; Yang, K.; Liu, S.; Hou, J.; and Yi, S. 2021a. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13668–13677.

Li, S.; Zhang, F.; Yang, K.; Liu, L.; Liu, S.; Hou, J.; and Yi, S. 2022a. Probing visual-audio representation for video highlight detection via hard-pairs guided contrastive learning. *arXiv preprint arXiv:2206.10157*.

Li, S.; Zhang, F.; Zhao, R.-W.; Feng, R.; Yang, K.; Liu, L.; and Hou, J. 2022b. Pyramid Region-based Slot Attention Network for Temporal Action Proposal Generation. *arXiv preprint arXiv:2206.10095*.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020a. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020b. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2: 429–450.

Li, W.; Milletarì, F.; Xu, D.; Rieke, N.; Hancox, J.; Zhu, W.; Baust, M.; Cheng, Y.; Ourselin, S.; Cardoso, M. J.; et al. 2019. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 133–141. Springer.

Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; and Dou, Q. 2021b. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*.

Liang, X.; Lin, Y.; Fu, H.; Zhu, L.; and Li, X. 2022. RSCFed: Random Sampling Consensus Federated Semi-supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10154–10163.

Lin, H.; Lou, J.; Xiong, L.; and Shahabi, C. 2021. Semifed: Semi-supervised federated learning with consistency and pseudo-labeling. *arXiv preprint arXiv:2108.09412*.

Liu, H.; Simonyan, K.; and Yang, Y. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

Liu, Q.; Yang, H.; Dou, Q.; and Heng, P.-A. 2021. Federated semi-supervised medical image classification via inter-client relation matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 325–335. Springer.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Niu, C.; Wu, F.; Tang, S.; Hua, L.; Jia, R.; Lv, C.; Wu, Z.; and Chen, G. 2020. Billion-scale federated learning on mobile clients: A submodel design with tunable privacy. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 1–14.

Pham, H.; Dai, Z.; Xie, Q.; and Le, Q. V. 2021. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11557–11568.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30: 1195–1204.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; and Khazaeni, Y. 2020. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*.

Wang, Q.; Bai, S.; Gao, J.; Yuan, Y.; and Li, X. 2021. Unsupervised Domain Adaptive Learning via Synthetic Data for Person Re-identification. *arXiv preprint arXiv:2109.05542*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Neurips*, 33: 6256–6268.

Xu, J.; Wang, S.; Wang, L.; and Yao, A. C.-C. 2021. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*.

Yang, H.; Zhu, S.; King, I.; and Lyu, M. R. 2011. Can Irrelevant Data Help Semi-Supervised Learning, Why and How? In *The 20th ACM International Conference on Information and Knowledge Management*.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2): 1–19.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021a. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419.

Zhang, Z.; Yang, Y.; Yao, Z.; Yan, Y.; Gonzalez, J. E.; Ramchandran, K.; and Mahoney, M. W. 2021b. Improving semi-supervised federated learning by reducing the gradient diversity of models. In *2021 IEEE International Conference on Big Data (Big Data)*, 1214–1225. IEEE.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Zhuang, W.; Wen, Y.; and Zhang, S. 2022. Divergence-aware Federated Self-Supervised Learning. In *International Conference on Learning Representations*.

Zou, Y.; Zhang, Z.; Zhang, H.; Li, C.-L.; Bian, X.; Huang, J.-B.; and Pfister, T. 2022. Pseudoseg: Designing pseudo labels for semantic segmentation. *International Conference on Learning Representations*.