Progressive Distribution Matching for Federated Semi-Supervised Learning

Dongping Liao*[1], Xitong Gao*[2,3], Yabo Xu[4], Chengzhong Xu†[1]

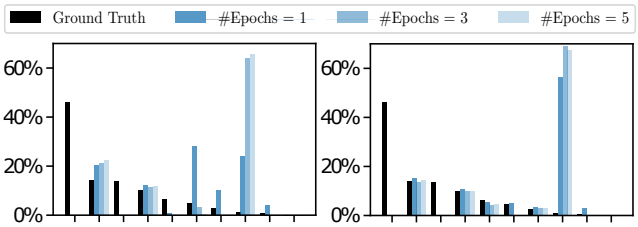
[1] State Key Lab of IoTSC, Department of Computer and Information Science, University of Macau, Macau SAR, China
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[3] Shenzhen University of Advanced Technology, Shenzhen, China
[4] DataStory Information Technology Co., Ltd
yb97428@um.edu.mo, xt.gao@siat.ac.cn, arber@datastory.com.cn, czxu@um.edu.mo

Abstract

Federated Learning (FL) enables collaborative learning from distributed data while preserving the privacy of participating clients. While supervised federated learning with labeled data has made notable strides and achieved success, federated semi-supervised learning (FSSL) lags in its progress. Existing works for FSSL heavily rely on fully-labeled clients, while ignoring the distribution of pseudo-labels generated from skewed unlabeled data. In this work, we offer empirical and theoretical insights into the challenges encountered when applying conventional semi-supervised algorithms in the federated regime. Specifically, we highlight how the inherent data heterogeneity in FSSL can exacerbate issues within the pseudo-labeling process. Motivated by these observations, we propose FL with progressive distribution matching (FedPDM) to regularize the distribution of pseudo-labels, aiming to progressively reshape it to align with the ground-truth distribution. The matching problem could be formulated as an optimal transport (OT) problem and efficiently solved by Sinkhorn-Knopp iteration. Through extensive experiments, we demonstrate the superiority of FedPDM on a variety of models and datasets compared with prior arts for FSSL.

1 Introduction

Federated Learning (FL) enables a group of clients to jointly optimize a model using distributed data, ensuring that the data remain locally stored and only privately accessible. This new paradigm distinguishes itself from traditional distributed optimization and offers advantages such as privacy preservation, secure aggregation, and improved communication efficiency. To date, there has been a notable surge of interest on this topic, such as pioneering applications in mobile computing (Hard et al. 2018) and healthcare domains (Liu et al. 2021). Despite extensive research efforts devoted to address data (Wang et al. 2020; Acar et al. 2020; Liao, Gao, and Xu 2024) and system heterogeneity (Diao, Ding, and Tarokh 2020; Liao et al. 2023; Gao et al. 2019), communication footprint (Wang, Lin, and Chen 2022; Hönig, Zhao, and Mullins 2022), these endeavors primarily assume the availability of fully-labeled data. However, real-world data are less frequently annotated, primarily due to constraints related to the budgets and expertise, and a lack of incentive for data owners to undertake such efforts. Consequently, demanding fully-labeled data for FL may prove impractical in many real-world scenarios. These constraints have thus propelled research into Federated Semi-Supervised Learning (FSSL) (Jeong et al. 2020; Liu et al. 2021).



(a) After 50 rounds. (b) After 200 rounds.

Figure 1: Classes (x-axis, sorted by true frequency) *vs.* Probability (y-axis, %) on CIFAR-10. As training rounds progress, client model gains an increasing prediction bias towards certain classes, leading to a distribution mismatch (deviation from "Ground-Truth") for unlabeled data. "Ground Truth": true class distribution, "#Epochs": number of local training epochs.

To leverage unlabeled data, earlier attempts for FSSL (Jeong et al. 2020; Liu et al. 2021; Yang et al. 2021) focus on adapting existing semi-supervised algorithms into federated scenarios. Nonetheless, these efforts tend to neglect the heterogeneity (*i.e.* non-IID-ness) inherent in local data. To resolve this issue, RSCFed (Liang et al. 2022) employs random sub-sampling and distance-reweighed model aggregation, while CBAFed (Li, Li, and Wang 2023) introduces adaptive confidence thresholding. Despite yielding positive outcomes, these approaches exhibit profound limitations. On one hand, they fail to uncover the underlying difficulties in accurately assigning pseudo-labels to samples — a task complicated by the presence of data heterogeneity. On the other hand, even with accurately assigned pseudo-labels, training still contends with the challenges of data heterogeneity, as they do not account for the data imbalance.

To this end, we delve into the distribution of the model

*These authors contributed equally to this work.
†Corresponding author.

predictions on unlabeled samples under the federated setting with aggressive data heterogeneity (see Appendix D for detailed setup). We start with exploratory experiments that incorporate FixMatch (Sohn et al. 2020), a strong centralized semi-supervised learning baseline. We visualize the averaged predictions of exponential moving averaged (EMA) model of a local client on unlabeled data along different local training epochs. The findings, shown in Figure 1, reveal two key aspects about its predictions: (1) the predictions exhibit a strong bias towards incorrect classes, suggesting a **distribution mismatch** even at the early stages of training; (2) the bias becomes more pronounced as training progresses, which shows that the model is **increasingly confident in its incorrect predictions**.

These observations shed light on the challenges encountered with incorporating existing centralized baselines into FL. The increasing confidence of the local model in incorrect classes worsens the distribution mismatch, compounding the effectiveness of pseudo-labeling and reinforcing the model's preference for these classes in a vicious cycle. Attempts to mitigate the effect, as explored in prior works (CBAFed, RSCFed, *etc.*) involve employing a global model updated per round for pseudo-labeling. However, the distribution mismatch problem persists in them, primarily due to the susceptibility of the global model to bias induced by local data heterogeneity, which are not addressed explicitly and adequately. Moreover, the scarcity of some classes in the local data coupled with dynamic changes in pseudo-label assignments, transforms FSSL into an even more formidable problem, characterized by the need to learn from a continuously shifting distribution. This exacerbates the problem of catastrophic forgetting (De Lange et al. 2021).

To address these challenges, we propose federated learning with progressive distribution matching (FedPDM). Ideally, the example in Figure 1 shows that the predicted class distribution of a high-quality model should thus closely match the "Ground Truth" (GT) label distribution. Deviations from the GT signifies a distribution mismatch, and leaving room for the pseudo-labels assignment to improve. The core principle of FedPDM is to refine the skewed distribution of local model predictions to align with an estimated distribution, which serves as proxy for the true class prior.

FedPDM formulates this distribution matching as an optimal transport problem (Peyré, Cuturi et al. 2017), adhering to the prior distribution constraint. This constrained problem can be efficiently optimized with Sinkhorn-Knopp iteration (Knight 2008). Distribution matching enjoys threefold advantages. First, it encourages the selection of underrepresented unlabeled samples, mitigating the risk of the model over-prioritizing incorrect classes. Second, it stabilizes the training dynamics of local client models on unlabeled data. Finally, an attractive feature of FedPDM is its seamless integration with other class-balancing techniques (Menon et al. 2020), providing an explicit means to tackle the data heterogeneity inherent in unlabeled data in FL.

To summarize, our contributions are as follows:

- **Problem.** We reveal the key challenge of distribution mismatch in existing FSSL algorithms, which provides a clear interpretation of the observed compromised performance in these scenarios.
- **Method.** We introduce FedPDM to progressively refine the distribution of local predictions, explicitly addressing the data heterogeneity inherent in unlabeled data.
- **Analysis.** We conduct a generalization analysis for FSSL through the lens of domain-adaptation theory. The theoretical analysis endorses our method design and offers insights about the effectiveness of FedPDM.
- **Performance.** As evinced by extensive experiments, FedPDM yields much improved pseudo-labels for unlabeled data. It outperforms existing best competitors by a clear margin, especially under high data heterogeneity.

2 Related Work

Federated Learning A seminal work on federated learning, FedAvg (McMahan et al. 2016), has ignited notable research interest in the paradigm of learning from large scale privacy-sensitive data in a distributed manner. A major challenge in this domain is the data heterogeneity (McMahan et al. 2016), stemming from localized private data, a significant impediment to the performance of FL (Hsieh et al. 2020). In response to this challenge, one set of methods advocates for the regularization of local updates. FedProx (Li et al. 2020) employs proximal regularization to constrain the local update in the vicinity of the global model. SCAFFOLD (Karimireddy et al. 2020) draws inspiration from variance reduction (Reddi et al. 2016) to correct local "client drift". Similarly, FedDyn (Acar et al. 2020) uses dynamic regularization to align global and local optimization objectives. Concurrently, another set of approaches aims to mitigate the deviation of client updates during the global aggregation stage. FedMA (Wang et al. 2019) utilizes the Bayesian non-parametric model to fusion local models to a global model with dynamically expanding capacity. Inspired by Bayesian uncertainty, FedBE (Chen and Chao 2020) samples a group of possible global models for ensemble distillation. Despite their progress, they are tailored for supervised learning from fully-labeled data, while we seek to exploit the unlabeled data to address label deficiency in FL.

Semi-supervised Learning presents a promising research direction to leverage the potential of unlabeled samples. Representative approaches include consistency regularization (Tarvainen and Valpola 2017) and pseudo-labeling (Rizve et al. 2020). **Consistency regularization** relies on the manifold (smoothness) assumption, typically instantiated through consistent constraints among various models (Rasmus et al. 2015) and data augmentations. This enables a model to learn meaningful representations by finding a smooth manifold. **Pseudo-labeling** involves selecting high-confidence predictions as pseudo-labels via entropy minimization (Grandvalet and Bengio 2004) or confidence thresholding (Lee et al. 2013). Hybrid methods (Berthelot et al. 2019) integrate and generalize various aspects of the above methods. For a more detailed taxonomy, please refer to (Yang et al. 2022). There has been a growing body of research (Wang et al. 2022) focusing on the learning dynamics of each class. However, these methods assume balanced datasets. In contrast, FSSL poses the challenge of learning

from unlabeled data with **unknown and skewed distribution**, presenting an unresolved issue we aim to address in this work.

Federated Semi-supervised Learning Early efforts on this topic focus on integrating semi-supervised methods into federated setting to exploit unlabeled data. FedConsist (Yang et al. 2021) applies consistency regularization on medical imaging data within the federated context. FedIRM (Liu et al. 2021) presents an inter-client relation matching scheme to communicate inherent relationship on diseases to other clients with limited labels. SemiFL (Diao, Ding, and Tarokh 2022) proposes alternate training to leverage clients with unlabeled data to boost the performance of a global model trained on a centralized labeled dataset. FEDLABEL (Cho, Joshi, and Dimitriadis 2023) selectively chooses the local or global model to pseudo-label unlabeled data. RSCFed (Liang et al. 2022) employs random consensus to obtain sub-consensus models, followed by a distance-reweighed model aggregation. However, its consistency regularization overlooks the local data heterogeneity. The supervision from the exponential moving-average (EMA) model could be easily affected by local dataset bias. CBAFed (Li, Li, and Wang 2023) designs class-balanced adaptive thresholds for each class and leverages pseudo-labeling to select training samples. The two categories above respectively represent consistency regularization and pseudo-labeling approaches in semi-supervised learning. Yet, they all failed to explicitly address local skewed data distribution, and thus still suffer from biased local training, even assuming unlabeled samples with *perfectly* assigned pseudo labels. In contrast, we propose a method to progressively match the distribution for each local client dataset, explicitly addressing the issue of skewed distribution.

Optimal Transport (Villani et al. 2009) is a mathematical framework that finds the most efficient way to transport mass from one distribution to another, minimizing total transportation cost while satisfying given constraints. As an effective tool to compare two probability distributions, OT has also gained prominence in computer vision under the name of Earth Mover's distance (Zhao, Yang, and Tao 2008). Recently, approximate solvers (Cuturi 2013) have enabled efficient scaling to large-scale problems, leading to its widespread adoption in diverse domains such as few-shot learning (Zhang et al. 2020), adversarial robustness (Bui et al. 2021), and learning from multi-modal data (Lee et al. 2019). Notably, Taherkhani *et al.* (Taherkhani et al. 2020) proposes to transport labels via hierarchical OT within the context of (centralized) semi-supervised learning. Nevertheless, several substantial obstacles prevent it from being easily applied to FSSL. First, it relies on a known distribution before training to adjust pseudo-label distribution, a luxury we lack in FL, requiring us to estimate throughout the training phase. Second, it operates under a balanced distribution, whereas our emphasis is on addressing the skewed data distributions inherent in local client datasets.

3 Preliminaries

We consider a classification problem as local training task with input space $\mathcal{X} = \mathbb{R}^d$ and output space $\mathcal{Y} = \{1, \dots, K\}$, where K is number of classes. Our goal is to learn a high-quality global model $f : \mathcal{X} \to \mathcal{Y}$.

Federated Semi-supervised Learning (FSSL) We consider two settings that represent the typical scenarios of FSSL. In the first (**Mixed**), we assume a group of clients $\{C_1, \dots, C_n\}$, and each client c possesses a mixture of labeled dataset $\mathcal{D}_c^l = \{(x_i^l, y_i^l)\}_{i=1}^{N_c^l}$ and unlabeled dataset $\mathcal{D}_c^u = \{(x_i^u)\}_{i=1}^{N_c^u}$. In the second (**Pure**), we assume m labeled clients $\{C_1^l, \dots, C_m^l\}$ and n unlabeled clients $\{C_{m+1}^u, \dots, C_{m+n}^u\}$, and their datasets can be defined accordingly. We introduce our proposed method under the first scenario, and the second could be similarly implemented by specializing different supervision losses for labeled and unlabeled clients. For the local training on labeled data, existing FSSL methods utilize the cross-entropy (CE) loss as the learning objective:

$$\mathcal{L}^l = \frac{1}{B} \sum_{b=1}^{B} \mathrm{H}\big(\boldsymbol{y}_b^l,\ \boldsymbol{p}_m(y_b^l \mid x_b^l)\big), \tag{1}$$

where B is the batch size, and $\mathrm{H}(\boldsymbol{u}, \boldsymbol{v})$ calculates cross-entropy between two probability distributions $\boldsymbol{u}$ and $\boldsymbol{v}$, and $\boldsymbol{p}_m$ represents the normalized prediction of a client model.

To leverage unlabeled data, pseudo-labeling methods generate artificial labels from unlabeled data by thresholding the output probability. This refers to the following loss function:

$$\mathcal{L}^u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbf{1}[\max(\boldsymbol{q}_b) \geq \tau]\, \mathrm{H}(\hat{\boldsymbol{q}}_b, \boldsymbol{q}_b), \tag{2}$$

where $\boldsymbol{q}_b = \boldsymbol{p}_m(y_b^u | x_b^u)$, and $\hat{\boldsymbol{q}}_b$ is a one-hot notation of $\arg\max(\boldsymbol{q}_b)$. The hyperparameter μ is a scaling factor that determines the batch size of unlabeled data, and τ is a threshold to control pseudo-labels selection.

Optimal Transport (OT) computes the minimal cost of transporting between probability measures. In this work, we focus on transport problem of discrete probability measures, and refer readers to (Peyré, Cuturi et al. 2017) for more general formulation. We start with two n- and m-dimensional discrete probability distributions $p = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $q = \sum_{j=1}^{m} b_i \delta_{y_j}$. Here, x_i and y_j denote locations, δ_{x_i} is the Dirac function located at x_i, implying a unit of mass which is infinitely concentrated at position x_i. The vectors $\boldsymbol{a} \in \Delta^n$ and $\boldsymbol{b} \in \Delta^m$ are two probability simplexes that define the weights at each location. The OT distance can be calculated by solving the constrained optimization problem:

$$d(p, q) = \min_{\boldsymbol{T} \in \Pi(p,q)} \{\langle \boldsymbol{T}, \boldsymbol{C} \rangle \triangleq \textstyle\sum_{i,j} \boldsymbol{T}_{ij} \boldsymbol{C}_{ij}\}, \tag{3}$$

where $d(p, q)$ is the weighted distance between p and q, and C_{ij} defines the transport cost matrix, $\langle \boldsymbol{T}, \boldsymbol{C} \rangle$ denotes the dot-product between the matrices $\boldsymbol{T}$ and $\boldsymbol{C}$, where $\boldsymbol{T}$ is the transport probability matrix subjects to $\Pi(p, q) \triangleq \{\boldsymbol{T} | \sum_{i=1}^{n} T_{ij} = b_j, \sum_{j=1}^{m} T_{ij} = a_i\}$, *i.e.*, the joint distribution of p and q. The target of OT is to find an optimal probability matrix $\boldsymbol{T}^\star$ (the optimal transportation plan) that minimizes the total transportation cost.

4 Progressive Distribution Matching

Figure 2 provides the high-level overview of the proposed FedPDM on a FSSL client. The heatmaps represent the output probabilities of batched data. For each batch of unlabeled data, we gather the data predictions from the EMA

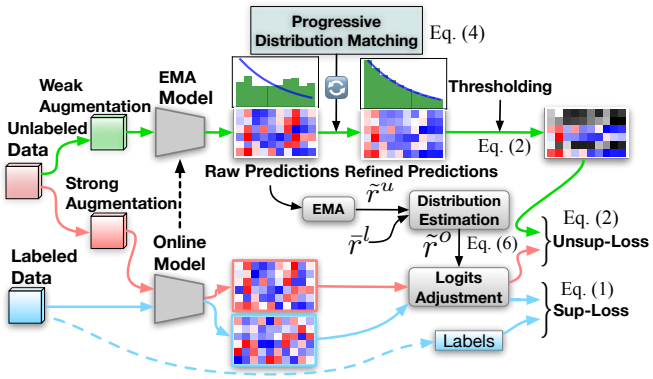


Figure 2: A high-level overview of proposed FedPDM on a federated semi-supervised learning client.

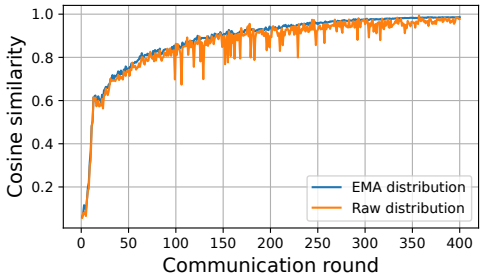


Figure 3: Comparing the cosine similarity between estimated distribution and ground-truth distribution of unlabeled data on CIFAR-100 over training rounds.

model with weak (random crop and flip) augmentations, and the online model with strong (RandAug (Cubuk et al. 2020)) augmentations. We refine them with progressive distribution matching to align with the estimated class prior distribution of unlabeled data ($\tilde{\boldsymbol{r}}^u$), which is obtained by EMA updates, and with the true distribution of labeled data ($\bar{r}^l$). The pseudo-labels are generated by thresholding the refined predictions and used as targets to optimize the online model with strongly augmented data. Finally, we address the local data heterogeneity with logits adjustment based on the estimated distribution ($\tilde{\boldsymbol{r}}^o$) for both supervised and unsupervised losses.

How to address distribution mismatch of pseudo labeling? The target of progressive distribution matching is to refine the skewed distribution of original EMA model predictions $\boldsymbol{P} = [\boldsymbol{p}_1, \dots, \boldsymbol{p}_N]^\top \in \mathbb{R}_+^{N\times K}$ based on a predefined cost matrix $\boldsymbol{C} \in \mathbb{R}_+^{N\times K}$ such that the refined pseudo-labels matrix $\boldsymbol{Q} = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_N]^\top \in \mathbb{R}_+^{N\times K}$ can closely reflect the true class prior distribution. Here, $\boldsymbol{Q}$ resembles the transportation plan $\boldsymbol{T}$ in Eq.3. We formulate the distribution matching as an optimal transport problem:

$$\begin{aligned} &\min_{\boldsymbol{Q}} \langle \boldsymbol{Q}, \boldsymbol{C}\rangle - \tfrac{1}{\lambda}\mathrm{H}(\boldsymbol{Q}) = -\textstyle\sum_{i=1}^{N}\sum_{j=1}^{K} Q_{ij}\log(P_{ij}Q_{ij}^{\frac{1}{\lambda}}) \\ &\quad s.t.\, Q_{ij} \geq 0,\ \boldsymbol{Q}^\top \mathbf{1}_N = \boldsymbol{r},\ \boldsymbol{Q}\mathbf{1}_K = \boldsymbol{c}. \end{aligned} \tag{4}$$

Here, we define $C_{ij} = -\log P_{ij}$, which implies that we assume lower transportation cost for predictions with higher confidence. Inspired by (Cuturi 2013), we also append an entropy regularization term with a trade-off hyperparameter λ to prevent the solution of $\boldsymbol{Q}$ being overly sparse, which can also be regarded as a smoothing regularization. This regularization also notably accelerates the computation of OT problem (Cuturi 2013). The first constraint in (4) limits the elements in $\boldsymbol{Q}$ to be non-negative, and the remaining two respectively define the desired property of row- and column-wise distribution of matrix $\boldsymbol{Q}$. The simplex $\boldsymbol{r}$ is a K-dimension vector that defines the class distribution of unlabeled data. Figure 3 demonstrates that while the distribution at each training round can be noisy and deviate from the true class prior. An EMA update can converge smoothly and yield high-quality approximation, thanks to the temporal ensemble nature of EMA, which leverages all past predictions.

The N-dimensional vector $\boldsymbol{c} = \mathbf{1}_N$ indicates each row of matrix $\boldsymbol{Q}$ is normalized to represent a probability distribution. This is also help to counteract the biased pseudo-label process. The optimal transport problem can be efficiently solved with Sinkhorn-Knopp algorithm (Cuturi 2013) by alternatively normalizing the row and column with scaling coefficient vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to satisfy the constraints. To achieve this, we use $\boldsymbol{M} \in \mathbb{R}_+^{N\times K}$ where $M_{ij} = e^{-\lambda C_{ij}} = P_{ij}^{\lambda}$, and optimize $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ iteratively by coordinate descent (Boyd and Vandenberghe 2004):

$$\boldsymbol{\alpha} \leftarrow \log \boldsymbol{c} - \log(\boldsymbol{M}e^{\boldsymbol{\beta}}),\ \boldsymbol{\beta} \leftarrow \log \boldsymbol{r} - \log(\boldsymbol{M}^\top e^{\boldsymbol{\alpha}}). \tag{5}$$

The refined predicted distribution of unlabeled data is thus $\boldsymbol{Q} = \mathrm{diag}(e^{\boldsymbol{\alpha}})\boldsymbol{M}\,\mathrm{diag}(e^{\boldsymbol{\beta}})$. In our implementations, we initialize a matrix $\boldsymbol{M}$ to store the predictions of unlabeled data of each client and update $\boldsymbol{M}$ progressively in each local optimization step. We apply distribution matching for each mini-batch of unlabeled data and refine predictions with $\boldsymbol{Q}$. The distribution matching operates locally on each client and incurs no additional information exchange among clients. We present the details in Algorithm 1 and defer the derivation to Appendix C.

It is worth noting that the estimated class prior distribution of unlabeled data plays a *pivotal* role in connecting the two important issues in FSSL, *i.e.*, the skewed pseudo-labeling problem and biased local training. On one hand, the estimated class prior distribution addresses pseudo-labeling distribution mismatch. On the other hand, we employ it to tackle the biased local training to stabilize the training dynamics of FSSL. Benefiting from this, FedPDM can seamlessly incorporate many class-balancing tools that leverage the class distribution of training data, granting us the flexibility and possibility to further push the frontier of FSSL algorithms. To our knowledge, this has not been investigated and exploited in existing FSSL works as the distribution of unlabeled data is not available in these works.

How to effectively tackle biased training on unlabeled data? Favorably, the estimated distribution $\boldsymbol{r}$ can be seamlessly incorporated to existing class-balancing techniques such as resampling or reweighing. This gives rise to debiased local training of clients on unlabeled data. Nevertheless, resampling methods incurs additional computation cost

Algorithm 1: FedPDM: Progressive Distribution Matching

ServerUpdate $(\mathbb{C}, \mathcal{R}, T)$**:**
- **for** each communication round $t = 1, 2, \dots T$ **do**
  - $S_t \leftarrow \texttt{sample}(\mathbb{C}, \mathcal{R})$ `// Client sampling`
  - **for** each client $c \in S_t$ **in parallel do**
    - $W_c^t \leftarrow$ **ClientUpdate**$(\mathcal{D}_c^l, \mathcal{D}_c^u, W_g^{t-1})$
  - $W_g^t = \sum_c^{|S_t|} \rho_c \cdot W_c^t$ `// Global aggregation`
- **return** W_g^T

ClientUpdate $(\mathcal{D}_c^l, \mathcal{D}_c^u, W_c^t)$**:**
- $\tilde{W}_c^t \leftarrow W_c^t$ `// Initialize EMA model`
- **for** each local epoch $e = 1, 2, \dots E$ **do**
  - **for** batch $(x_B^l, y_B^l), (x_{\mu B}^u, y_{\mu B}^u) \in \mathcal{D}_c^l, \mathcal{D}_c^u$, **do**
    - $\mathcal{L}^l \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathrm{H}\big(\boldsymbol{y}_b^l,\ \boldsymbol{p}_m(y_b^l \mid x_b^l)\big)$
    - $\boldsymbol{P}_{\mu B} \leftarrow f_{\tilde{W}_c^t}(x_{\mu B}^u)$ `// Batch predictions`
    - $\boldsymbol{Q}_{\mu B} \leftarrow$ **Sinkhorn** $(\boldsymbol{P}_{\mu B}, \boldsymbol{r}, \boldsymbol{c})$
    - $\mathcal{L}^u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbf{1}[\max(\boldsymbol{q}_b) \geq \tau] \mathrm{H}(\hat{\boldsymbol{q}}_b, \boldsymbol{q}_b)$
    - $\mathcal{L} = \mathcal{L}^l + \gamma \mathcal{L}^u$
    - $W_c^t \leftarrow W_c^t - \eta \nabla_{W_c^t} L$ `// Online model`
    - $\tilde{W}_c^t \leftarrow \beta_1 \tilde{W}_c^t + (1 - \beta_1) W_c^t$ `// EMA model`
    - $\boldsymbol{r} \leftarrow \beta_2 \boldsymbol{r} + (1 - \beta_2) \boldsymbol{P}_{\mu B}^\top \mathbf{1}_B$ `// EMA dist.`
- **return** $\tilde{W}_c^t$

Sinkhorn $(\boldsymbol{P}_{\mu B}, \boldsymbol{r}, \boldsymbol{c})$**:**
- $\boldsymbol{M} \leftarrow \texttt{Update}(\boldsymbol{M}, \boldsymbol{P}_{\mu B})$ `// Update cost matrix`
- **for** each iteration $i = 1, \dots, I_m$ **do**
  - `// Sinkhorn-Knopp Iteration`
  - $\boldsymbol{\alpha} \leftarrow \log \boldsymbol{c} - \log(\boldsymbol{M} e^{\boldsymbol{\beta}})$; $\boldsymbol{\beta} \leftarrow \log \boldsymbol{r} - \log(\boldsymbol{M}^\top e^{\boldsymbol{\alpha}})$
- $\boldsymbol{Q} = \mathrm{diag}(e^{\boldsymbol{\alpha}}) \boldsymbol{M} \, \mathrm{diag}(e^{\boldsymbol{\beta}})$ `// Refined preds`
- $\boldsymbol{Q}_{\mu B} \leftarrow \texttt{Extract}(\boldsymbol{Q}, \texttt{Index}(\boldsymbol{P}_{\mu B}))$
- **return** $\boldsymbol{Q}_{\mu B}$

in FL local training, and the missing classes of local dataset can not be over-sampled. FedPDM thus adopts *logit adjustment* (Menon et al. 2021), a lightweight reweighing remedy that operates only on model predictions without neural overhead. This gives rise to the following loss function:

$$\mathcal{L}^{\mathrm{LA}}(\mathbf{p}, y) = -\log \frac{\exp\left(p_y + \kappa \log r_y\right)}{\sum_{j=1}^{K} \exp\left(p_j + \kappa \log r_j\right)}, \tag{6}$$

where κ is hyperparameter to control the strength of logits adjustment, and $\boldsymbol{r} = [r_1, \dots, r_K]$ is an estimated class distribution. As clients possess a mixture of labeled and unlabeled data, we first calculate the overall distribution of local labeled and unlabeled data and then use it for logits adjustment of all samples. The adjustment can be achieved by replacing the cross-entropy loss in Algorithm 1 with the loss defined in (6). Since the estimated distribution of unlabeled data can be inaccurate at the beginning of local training, we apply a linear-rampup schedule to gradually increase the impact of logits adjustment.

We note that existing FSSL methods lack such consideration and they thus still suffer from the dreaded distribution shift in semi-supervised *local* training despite their innovations. Note that distribution matching and the class-balancing techniques can compliment each other jointly. This is due to the former is conducive to reshaping the predictions to better align with the true class prior. Benefiting from this alignment, the class-balancing techniques can be applied on these refined predictions more efficiently than on a skewed counterpart. This is also supported by the analysis in Theorem 1, and ablation studies in Appendix H.2.

Since the thresholding operator in pseudo-labeling process causes varying number of selected unlabeled data at each communication round, we introduce an adaptive averaging method that considers the effective number of training data. We define a weight ratio for each client c:

$$\rho_c = (\tilde{N}_c^u + N_c^l) \,/\, \textstyle\sum_{i=1}^{|S_t|} (\tilde{N}_i^u + N_i^l), \tag{7}$$

where $\tilde{N}_i^u$ is the selected total number of unlabeled data of client i. The adaptive aggregation scheme automates the evaluation of client aggregation weights. It specifically frees us from manually tuning the aggregation weights of labeled and unlabeled clients as in RSCFed (Liang et al. 2022).

5 Theoretical Analysis

To gain a deeper understanding about the intrinsic difficulty of FSSL under heterogeneous data, we conduct the generalization analysis. Following the motivation from multi-source domain adaptation (Mansour, Mohri, and Rostamizadeh 2008), we decompose the domain gap between the data distribution of local client and global distribution into two parts, namely, label distribution shift represented by local domain M_k and label noise denoted by N_k. The former focuses on the imbalance of data quantity of each class, and the latter is specialized for pseudo-labeling error on unlabeled data in semi-supervised learning. Below we consider a FSSL system with K clients. For an in-depth proof, please refer to Appendix B.

Theorem 1 (Generalization Bound for FSSL)**.** *Let G be the global domain, and the risk of a hypothesis h on G is defined as $\mathcal{L}_G(h)$. Assume h_k is the hypothesis learned on local domain for client k and $h = \frac{1}{K} \sum_{k=1}^{K} h_k$ is the hypothesis of global domain. Let $\tilde{\mathcal{D}}_G$ and its parallel notations $\tilde{\mathcal{D}}_{M_k}$ and $\tilde{\mathcal{D}}_{N_k}$ be the induced distributions over the feature space $\mathcal{Z}$. With probability at least $1 - \delta$:*

$$\begin{aligned}
\mathcal{L}_G(h) \leq\ & \tfrac{1}{K} \textstyle\sum_{k \in [K]} [\hat{\mathcal{L}}_{M_k}(h_k) + \hat{\mathcal{L}}_{N_k}(h_k)] \\
&+ \tfrac{1}{K} \textstyle\sum_{k \in [K]} [d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{\mathcal{D}}_{M_k}, \tilde{\mathcal{D}}_G) + d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{\mathcal{D}}_{N_k}, \tilde{\mathcal{D}}_G)] \\
&+ \tfrac{\lambda_k}{K} + 4 \sqrt{\tfrac{1}{m} (d \log \tfrac{2em}{d} + \log \tfrac{4K}{\delta})},
\end{aligned} \tag{8}$$

where $\hat{\mathcal{L}}_{M_k}(h_k)$ and $\hat{\mathcal{L}}_{N_k}(h_k)$ are the empirical risks of h_k, $d_{\mathcal{H}\triangle\mathcal{H}}(\cdot, \cdot)$ measures the $\mathcal{H}$-divergence between two distributions, $\lambda_k := \min_h \mathcal{L}_G(h) + \mathcal{L}_{M_k}(h) + \mathcal{L}_{N_k}(h)$ represents the minimum of the combined risks on domains G, M_k and N_k, and m is the number of samples on each local clients.

Remark 1. If the unlabeled training samples are precisely annotated in pseudo-labeling process, Theorem 1 degenerates to the generalization bound of FedAvg (McMahan et al. 2016) on *fully-labeled* data.

What the generalization bound reveals? Theorem 1 indicates the generalization upper-bound of global hypothesis is primarily contingent on three terms: **(1)** the averaged empirical risk of local hypothesis on skewed domains, **(2)** the discrepancy between local and global distribution raised by *local distribution shift* and *pseudo-labeling error*, and **(3)** the number of local samples and the complexity of hypothesis. Notably, the above analysis also suggests that the generalization performance of existing FSSL algorithms (Liu et al. 2021; Liang et al. 2022; Li, Li, and Wang 2023) is upper-bounded by FedAvg (McMahan et al. 2016) on fully-labeled data, as they do not account for the data heterogeneity issue explicitly. In contrast, FedPDM takes a step forward to address the distribution mismatch to minimize the domain gaps. Specifically, our distribution matching intends to minimize $d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{\mathcal{D}}_{N_k}, \tilde{\mathcal{D}}_G)$ incurred by pseudo-labeling error.

Why debiased local training works? Appendix B.3 presents additional insights of how debiased local training (6) addresses the label distribution shift, which contributes to reducing $d_{\mathcal{H}\triangle\mathcal{H}}(\tilde{\mathcal{D}}_{M_k}, \tilde{\mathcal{D}}_G)$ of generalization bound (8).

6 Evaluation of FSSL Algorithms

Benchmark datasets. We conduct experiments on five prevalent datasets for FSSL: Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), SVHN (Netzer et al. 2011), CIFAR-10 (Krizhevsky, Hinton et al. 2009), CIFAR-100 (Krizhevsky, Hinton et al. 2009), and ISIC2018 (Codella et al. 2019). The details for dataset processing and configurations are outlined in Appendix E.1.

FSSL scenarios. Following (Jeong et al. 2020; Li, Li, and Wang 2023; Liang et al. 2022), we assume the labeled data availability on clients rather than server. We consider two FSSL scenarios, which we refer to as **Mixed** and **Pure** respectively. For the **Mixed** scenario, we simulate a mixture of labeled and unlabeled datasets for each client. To achieve this, we randomly split a subset from the original dataset as the labeled dataset and construct the remaining data as the unlabeled dataset. We partition the labeled and unlabeled datasets separately to multiple sub-datasets according to Dirichlet distribution $Dir(\alpha)$, where α controls the heterogeneity degree of local clients. A small α indicates high data heterogeneity, while $\alpha \to \infty$ gives class-balanced data partitions. Note that the class prior distributions of labeled and unlabeled datasets are different as exemplified in Figure 9 of Appendix E.1, and it is thus not feasible to infer the distribution of unlabeled data from labeled data. For the **Pure** scenario, we directly partition the original dataset into multiple heterogeneous subsets. We set 10 local clients for each scenario. Specifically, we have 1 labeled and 9 unlabeled clients for the latter scenario to align with existing works (Liang et al. 2022; Li, Li, and Wang 2023). For all experiments, we report the best accuracy of the global model obtained by each method.

Training setups. For all experiments, we used the SGD optimizer with batch size 64, momentum 0.9, and weight decay $5e^{-4}$. We utilized ResNet18 (He et al. 2016) for main evaluation. We set the number of communication rounds T as 400 and local epoch e as 5. For **Mixed**, we set the learning rate as 0.03. For **Pure**, the learning rate for labeled and unlabeled clients are respectively set as 0.03 and 0.02. A cosine learning rate scheduler with the total decay rounds being set as 400. This aligns with experimental settings of existing semi-supervised literature (Sohn et al. 2020). To select high-quality pseudo-labels, we set the confidence threshold $\tau = 0.95$ on all datasets without further tuning. After each local training round, the clients transmit the EMA model to the server. Appendix E.2 provides the remaining hyperparameters.

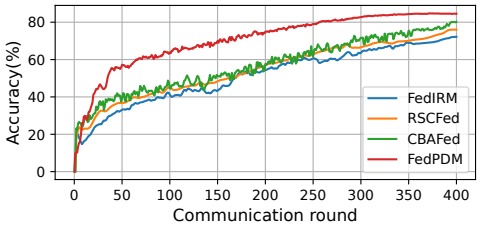


Figure 4: Convergence curve comparisons of FSSL methods (rounds *vs.* accuracy (%)) on CIFAR-10. FedPDM enjoys much faster convergence and much higher final accuracy than the competing FSSL methods. For a similar comparison on CIFAR-100, please refer to Figure 10 in Appendix H.1.

Compared methods. We compare with prior arts for FSSL: (1) **FedIRM** (Liu et al. 2021), which adds the inter-class relation loss computed with predictions of labeled clients to unlabeled clients; (2) **RSCFed** (Liang et al. 2022), which leverages random sub-consense sampling to optimize local clients and uses distance-aware aggregation to obtain global model; (3) **CBAFed** (Li, Li, and Wang 2023), which introduces adaptive threshold to improve pseudo-labeling. Following RSCFed and CBAFed for referential baselines: for **Pure**, we also included the accuracy of FedAvg (McMahan et al. 2016) trained with 1 labeled client as a lower bound, and FedAvg (McMahan et al. 2016) trained with 10 labeled client as an upper bound; for **Mixed**, we used FedAvg (McMahan et al. 2016) trained on labeled dataset.

FedPDM achieves better converged accuracies. We present the main evaluation of FedPDM and compared methods under two FSSL scenarios in Table 1. We observed FedPDM consistently yields the highest accuracies compared with other FSSL algorithms. On the CIFAR-100 dataset under **Mixed** scenario, it remarkably surpasses the best competitor by 4.7%. Importantly, it also indicates that the advantage can be generalized on large-scale datasets with higher image resolution such as ISIC2018, which extends its applicability to real-world tasks. Figure 4 illustrates the convergence curves of evaluated FSSL methods under **Mixed**. It can be observed that FedPDM converges notably faster than its competitors. Benefiting from the regularization effect introduced by distribution matching, it also manifests stable and smooth convergence behavior.

FedPDM thrives under aggressive data heterogeneity. In addition to the main evaluation, we were also interested

Scenario	Supervision	Method	FMNIST	SVHN	CIFAR-10	CIFAR-100	ISIC2018
Mixed	Fully-Supervised	If all clients have labels (Upper bound)	93.27	96.17	94.18	78.26	85.81
		If all clients trained with labeled data only (Lower bound)	73.52	71.95	66.96	63.32	66.20
	Semi-Supervised	FEDIRM (Liu et al. 2021)	82.45	83.49	72.19	63.73	66.71
		RSCFED (Liang et al. 2022)	82.88	84.25	76.10	67.22	68.59
		CBAFED (Li, Li, and Wang 2023)	83.62	83.93	80.20	67.57	68.42
		FEDPDM (ours)	**85.19**	**87.69**	**83.74**	**72.14**	**69.50**
Pure	Fully-Supervised	If all clients have labels (Upper bound)	93.27	96.17	94.18	78.26	85.81
		If trained with the only 1 labeled client (Lower bound)	71.79	86.72	63.59	36.84	69.20
	Semi-Supervised	FEDIRM (Liu et al. 2021)	76.35	85.42	64.16	40.23	70.20
		RSCFED (Liang et al. 2022)	78.83	87.41	66.92	41.59	71.54
		CBAFED (Li, Li, and Wang 2023)	81.49	89.90	69.65	43.66	72.05
		FEDPDM (ours)	**85.76**	**94.21**	**74.02**	**47.71**	**73.89**

Table 1: A performance overview of compared methods. We evaluate all methods under **Mixed** and **Pure** FSSL scenarios. **Mixed** trains 10 clients with a mix of labeled and unlabeled data, and **Pure** trains 10 clients, where only 1 has labeled data.

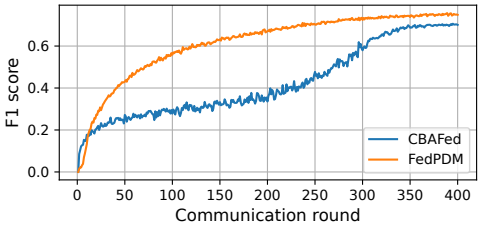


Figure 5: Comparing the F1-score of pseudo-labels (rounds *vs.* GT) on the CIFAR-100 dataset.

Method	CIFAR-10	CIFAR-100	ISIC2018
Upper bound	90.14	75.39	81.26
Lower bound	62.73	59.25	61.43
FEDIRM	68.55	58.01	61.60
RSCFED	71.48	61.41	62.83
CBAFED	72.61	63.60	63.02
FEDPDM (ours)	**78.96**	**69.40**	**67.54**

Table 2: Comparison of performance for **Mixed** under aggressive data heterogeneity with $\alpha = 0.1$ on 3 datasets.

in how pushing the degree of data heterogeneity could affect the performance of FSSL algorithms. We investigated a more aggressive heterogeneous data partition with $\alpha = 0.1$. Table 2 presents the results on three datasets. Unlike other FSSL algorithms, We highlight that FedPDM still thrives at high degrees of data heterogeneity, gaining a clear advantage over compared methods. In contrast, existing algorithms are oblivious to local training bias on unlabeled samples, presumably due to the lack of class distribution of unlabeled datasets. Their performance could thus be compromised easily under more difficult FSSL settings if the local bias is not properly addressed. Remarkably, on the large-scale real-world dataset ISIC2018, FedPDM only manifests a slight 1.96% performance drop when the data heterogeneity increases, whereas FedIRM suffers from 5.11% accuracy degradation. This also suggests that distribution matching and debiased local training for unlabeled data can push the limits of FSSL to more challenging scenarios.

FedPDM improves the quality of pseudo-labels. As pseudo-labels have a direct impact on the learning dynamics of local clients, we compared the quality of pseudo-labels between FedPDM and CBAFed. Since local unlabeled datasets are imbalanced, we used the multi-class F1-score to take into account both *precision* and *recall* metrics. In Figure 5, we highlight that FedPDM produces better pseudo-labels than CBAFed, and exhibits smooth and stable improvements as training progresses. This also explains the faster convergence of FedPDM over others.

Additional results and discussion. Table 7 in Appendix H reports the *ablations* to verify the impact of each component. To sum up, the distribution matching and debiased local training with logits adjustment jointly boost each other. We believe debiased training can maximize its effect when the distribution being applied best fits the GT distribution of the unlabeled dataset. This also suggests that debiasing FSSL training requires attention to the skewed distribution of the local model's predictions. We defer *hyperparameter sensitivity* to Appendix F, *computation overhead analysis* to Appendix G, and more experimental results and discussion to Appendix H.

7 Conclusion

We empirically reveal the pseudo-labeling of FSSL can be easily misled by unlabeled data heterogeneity. This causes distribution mismatch that hinders the correct assignment of pseudo-labels. We presented FedPDM, a novel FSSL method that aligns the unlabeled data predictions to their class prior distribution and tackles biased local training with an estimated class prior. Theoretical analysis revealed the limitations of existing FSSL algorithms and provides insights of our method design. Experiments showed that FedPDM can substantially improve the quality of pseudo-labels and even widen its lead over other approaches at aggressive data heterogeneity. We believe this paves the way for future work to address the label deficiency of FL algorithms, and consequently broaden their applicability on real-world tasks with limited annotations.

Acknowledgments

This work is supported in part by Science and Technology Development Fund of Macau SAR (Nos. 0081/2022/A2, 0123/2022/AFJ), National Natural Science Foundation of China (No. 62376263), Natural Science Foundation of Guangdong (No. 2024A1515030209), and Shenzhen Science and Technology Innovation Commission (No. JCYJ20230807140507015). This work was carried out in part at SICC, which is supported by SKL-IOTSC, University of Macau.

References

Acar, D. A. E.; Zhao, Y.; Matas, R.; Mattina, M.; Whatmough, P.; and Saligrama, V. 2020. Federated Learning Based on Dynamic Regularization. In *International Conference on Learning Representations*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Boyd, S. P.; and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Bui, A. T.; Le, T.; Tran, Q. H.; Zhao, H.; and Phung, D. 2021. A Unified Wasserstein Distributional Robustness Framework for Adversarial Training. In *International Conference on Learning Representations*.

Chen, H.-Y.; and Chao, W.-L. 2020. FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning. In *International Conference on Learning Representations*.

Cho, Y. J.; Joshi, G.; and Dimitriadis, D. 2023. Local or Global: Selective Knowledge Assimilation for Federated Learning with Limited Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17087–17096.

Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M. E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7): 3366–3385.

Diao, E.; Ding, J.; and Tarokh, V. 2020. HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients. In *International Conference on Learning Representations*.

Diao, E.; Ding, J.; and Tarokh, V. 2022. SemiFL: Semi-supervised federated learning for unlabeled clients with alternate training. *Advances in Neural Information Processing Systems*, 35: 17871–17884.

Gao, X.; Zhao, Y.; Dudziak, Ł.; Mullins, R.; and Xu, C.-z. 2019. Dynamic Channel Pruning: Feature Boosting and Suppression. In *International Conference on Learning Representations*.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

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 (2018). *arXiv preprint arXiv:1811.03604*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hönig, R.; Zhao, Y.; and Mullins, R. 2022. DAdaQuant: Doubly-adaptive quantization for communication-efficient Federated Learning. In *International Conference on Machine Learning*, 8852–8866. PMLR.

Hsieh, K.; Phanishayee, A.; Mutlu, O.; and Gibbons, P. 2020. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, 4387–4398. PMLR.

Jeong, W.; Yoon, J.; Yang, E.; and Hwang, S. J. 2020. Federated Semi-Supervised Learning with Inter-Client Consistency & Disjoint Learning. In *International Conference on Learning Representations*.

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.

Knight, P. A. 2008. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1): 261–275.

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

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. Atlanta.

Lee, J.; Dabagia, M.; Dyer, E.; and Rozell, C. 2019. Hierarchical optimal transport for multimodal distribution alignment. *Advances in neural information processing systems*, 32.

Li, M.; Li, Q.; and Wang, Y. 2023. Class Balanced Adaptive Pseudo Labeling for Federated Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16292–16301.

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

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.

Liao, D.; Gao, X.; and Xu, C. 2024. Impartial Adversarial Distillation: Addressing Biased Data-Free Knowledge Distillation via Adaptive Constrained Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3342–3350.

Liao, D.; Gao, X.; Zhao, Y.; and Xu, C.-Z. 2023. Adaptive Channel Sparsity for Federated Learning Under System Heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20432–20441.

Liu, Q.; Yang, H.; Dou, Q.; and Heng, P.-A. 2021. Federated semi-supervised medical image classification via inter-client relation matching. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, 325–335. Springer.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2008. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21.

McMahan, H. B.; Moore, E.; Ramage, D.; and y Arcas, B. A. 2016. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*.

Menon, A. K.; Jayasumana, S.; Jain, H.; Veit, A.; Kumar, S.; and Rawat, A. S. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Peyré, G.; Cuturi, M.; et al. 2017. Computational optimal transport. *Center for Research in Economics and Statistics Working Papers*, (2017-86).

Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; and Raiko, T. 2015. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28.

Reddi, S. J.; Hefny, A.; Sra, S.; Póczos, B.; and Smola, A. 2016. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, 314–323. PMLR.

Rizve, M. N.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2020. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In *International Conference on Learning Representations*.

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.

Taherkhani, F.; Dabouei, A.; Soleymani, S.; Dawson, J.; and Nasrabadi, N. M. 2020. Transporting labels via hierarchical optimal transport for semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 509–526. Springer.

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.

Villani, C.; et al. 2009. *Optimal transport: old and new*, volume 338. Springer.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; and Khazaeni, Y. 2019. Federated Learning with Matched Averaging. In *International Conference on Learning Representations*.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33: 7611–7623.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; et al. 2022. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*.

Wang, Y.; Lin, L.; and Chen, J. 2022. Communication-efficient adaptive federated learning. In *International Conference on Machine Learning*, 22802–22838. PMLR.

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

Yang, D.; Xu, Z.; Li, W.; Myronenko, A.; Roth, H. R.; Harmon, S.; Xu, S.; Turkbey, B.; Turkbey, E.; Wang, X.; et al. 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. *Medical image analysis*, 70: 101992.

Yang, X.; Song, Z.; King, I.; and Xu, Z. 2022. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12203–12213.

Zhao, Q.; Yang, Z.; and Tao, H. 2008. Differential earth mover's distance with its applications to visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2): 274–287.