Dual-calibrated Co-training Framework for Personalized Federated Semi-Supervised Medical Image Segmentation

Delin Pan[1], Jiansong Fan[1], Jie Zhu[1], Lihua Li[2], Xiang Pan[1,3]*

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
[2]Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou, China
[3]The PRC Ministry of Education Engineering Research Center of Intelligent Technology for Healthcare, Wuxi, Jiangsu 214122, China
6233115017@stu.jiangnan.edu.cn, fan_jiansong@163.com, 6233110009@stu.jiangnan.edu.cn
lilh@hdu.edu.cn, xiangpan@jiangnan.edu.cn

Abstract

Federated Semi-Supervised Learning (FSSL) has emerged as a crucial topic in medical image analysis, allowing multiple medical institutions to collaboratively train a global model using limited labeled data. However, existing FSSL methods focus solely on an effective combination of federated learning and semi-supervised learning, ignoring the heterogeneity of client data and the inadaptability of semi-supervised methods in diverse environments, which leads to knowledge bias in local models and impedes stable convergence. To this end, we explore the application of personalization in FSSL and propose a novel dual-calibrated co-training framework. To adapt to the unique feature distribution of client data, we consider collaborative relationships among clients to aggregate a personalized model for each client. We further build a dual-student architecture with the personalized model and private local model on the client side, which encourages model disagreement for co-training while enhancing participant privacy. Most importantly, we design dual calibration strategies that adaptively optimize the model: Local calibration improves the boundary discrimination of the local model by dynamically replacing pseudo-label boundary patches; Global calibration corrects model direction based on the real-time perception of the biases between local dual-student models. Experimental results show the effectiveness of our method on a private medical dataset and two public medical datasets.

Code —
https://github.com/Medical-AI-Lab-of-JNU/PFSSL

Introduction

Federated Learning (FL) has garnered considerable attention recently in medical image analysis, which enables multiple medical institutions (i.e., clients) to train a shared global model without centralizing data (McMahan et al. 2017; Konečnỳ et al. 2016; Liu et al. 2020). By ensuring patient data privacy while meeting the necessity for collaborative learning in the healthcare field, FL provides a way for more precise diagnoses (Antunes et al. 2022). Despite the potential of FL, the requirement for a large amount of fully labeled data restricts its practicality. In real-world medical scenarios, it is difficult for medical institutions to provide enough medical image annotations due to constraints such as limited available expertise, time, and the high costs of data annotation. Obtaining pixel-level annotations on a large scale is undoubtedly a significant challenge, especially for segmentation tasks that require pixel-level detail. To handle this problem, many researchers have considered an effective combination of FL and semi-supervised learning. This innovative paradigm allows multiple clients to utilize limited labeled data and abundant unlabeled data to facilitate collaborative model training in a distributed environment (Jeong et al. 2020; Liang et al. 2022; Lin et al. 2021).

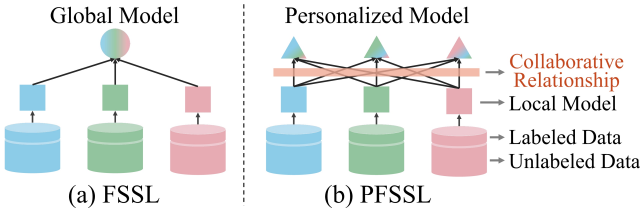


Figure 1: Illustration of the existing FSSL method and our PFSSL method. Existing methods simply combine semi-supervised methods with federated averages, while our method considers collaborative relationships and aggregates a personalized model for each client.

In previous attempts, federated semi-supervised learning (FSSL) has been extensively studied. FedIRM (Liu et al. 2021) addresses the deficiency of task knowledge at unlabeled clients by extracting useful information through inter-client relation matching. FedCD (Liu, Wu, and Qin 2024) introduced a class awareness balance module to explore balanced learning of rare classes in unlabeled clients. Although these methods have considered the inter-client relationships to facilitate knowledge transfer between labeled and unlabeled data, their performance still faces two significant challenges: data heterogeneity among multiple clients and the inadaptability of semi-supervised methods in distributed environments. As depicted in Figure 2, the main reasons for performance degradation in traditional FSSL methods are two-fold: 1) data heterogeneity may cause the aggregated model to absorb varied experiences, leading to catastrophic knowledge forgetting and repetitive local self-training pat-

*Corresponding author

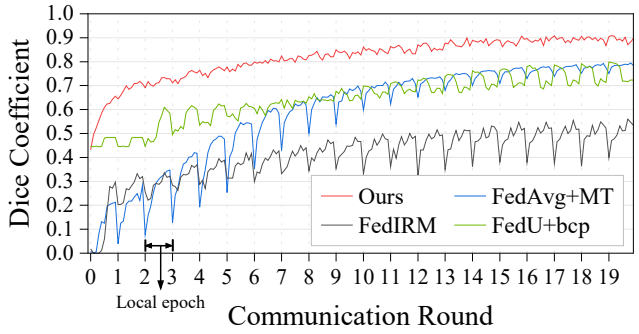


Figure 2: Traditional FSSL methods exhibit periodic fluctuations in test dice coefficient during training with federated aggregation rounds, struggling to adapt to data heterogeneity among clients. Our method is unaffected by aggregation and consistently delivers significantly better results than other methods. The experiment was based on a representative client using the Polyp dataset, with 10 local training epochs and 20 federated communication rounds. More details are described in the Section Experiments.

terns that struggle to converge stably; 2) the inadaptability of generalized semi-supervised methods with diverse data may result in significant noise in pseudo-labels, causing local models to learn incorrect decision boundaries and impeding high performance.

Hence, we propose a personalized federated semi-supervised learning (PFSSL) framework with co-training via dual calibration. As shown in Figure 1, we consider constructing rule relationship graphs of models to aggregate personalized models for clients, enabling each client to adapt to its unique data characteristics (Luo et al. 2021a). In this way, clients can gather preferred knowledge by matching with collaborative clients that have similar data distributions, which significantly alleviates the impact of data heterogeneity. To enhance client privacy protection (Kalra et al. 2023) and promote learning diversity, we establish a single-teacher-dual-student mechanism on the client side. The aggregated personalized model and the local private model serve as dual-student networks, with co-training between student networks to facilitate the integration of the teacher model (Hung et al. 2018; Yu et al. 2019; Chen et al. 2021). Crucially, we design dual calibration strategies (local calibration and global calibration) that flexibly update the local model based on the client's data distribution and learning state to alleviate local maladaptation. For local calibration, reliable knowledge with labeled boundaries is gathered in a patch-wise manner and utilized to dynamically replace pseudo-label boundary patches, thereby creating highly credible pseudo-labels. This process calibrates the ambiguous boundary semantics of unlabeled data, reducing the misleading impact of noise on model learning. The global calibration perceives the differences between the local model and the personalized model, using gradient updates to correct inconsistencies and guide the model in the right learning direction. Overall, our main contributions can be summarized as follows:

- We propose a novel FSSL method to address performance biases caused by heterogeneous data and the inadaptability of semi-supervised methods. Unlike existing FSSL methods, we incorporate personalized learning in FSSL.
- Our approach involves co-training for the local model and the personalized model aggregated by rule relationship graphs, especially designing dual calibration approaches to optimize local model performance and guide the model's learning direction, which achieves adaptive optimization for clients.
- We validated our method on liver organ, skin cancer lesion, and colorectal polyp segmentation tasks. Our method demonstrated remarkable segmentation results, showing the great potential of PFSSL.

Related Work

Personalized Federated Learning

Federated Learning (FL) is a collaborative learning framework that aims to train a global consensus model via coordinated communication with multiple clients. The data heterogeneity causes differences in client data distribution, making it challenging for a single global model to meet the diverse requirements of all clients. Consequently, Personalized Federated Learning (PFL) has emerged as an appropriate alternative solution (Kulkarni, Kulkarni, and Pant 2020; Tan et al. 2022). PFL can be viewed as a multi-task learning strategy, shifting from a traditional single server-centric learning task to multiple client-centric learning tasks (Marfoq et al. 2021). For instance, FedU (Dinh et al. 2022) proposed using Laplacian regularization to explicitly leverage the relationships among client models for multi-task learning. Pgfed (Luo et al. 2023) considers client risks to construct a personalized global objective for each client's local task. However, despite considering partial collaboration factors, these PFL methods lack a comprehensive and rigorous guideline for cooperation, which limits their adaptability to heterogeneous scenarios in the real world.

Semi-Supervised Learning

Semi-supervised learning (SSL) was proposed to improve supervised learning performance by utilizing a large amount of unlabeled information. Existing popular methods are mainly divided into pseudo-labeling and consistency regularization. Pseudo labeling (Lee et al. 2013) attempts to generate pseudo labels that are similar to the ground truth, and models are trained as in supervised learning. Consistency regularization (Lee et al. 2022; Tarvainen and Valpola 2017) enforces the model's outputs of different perturbed versions of the same input to be consistent. Current state-of-the-art approaches have incorporated these two strategies and shown satisfactory results for image segmentation tasks (Sohn et al. 2020; Zhang et al. 2021). Building on this foundation, we explored semi-supervised segmentation adaptive learning algorithms in the federated learning scenario.

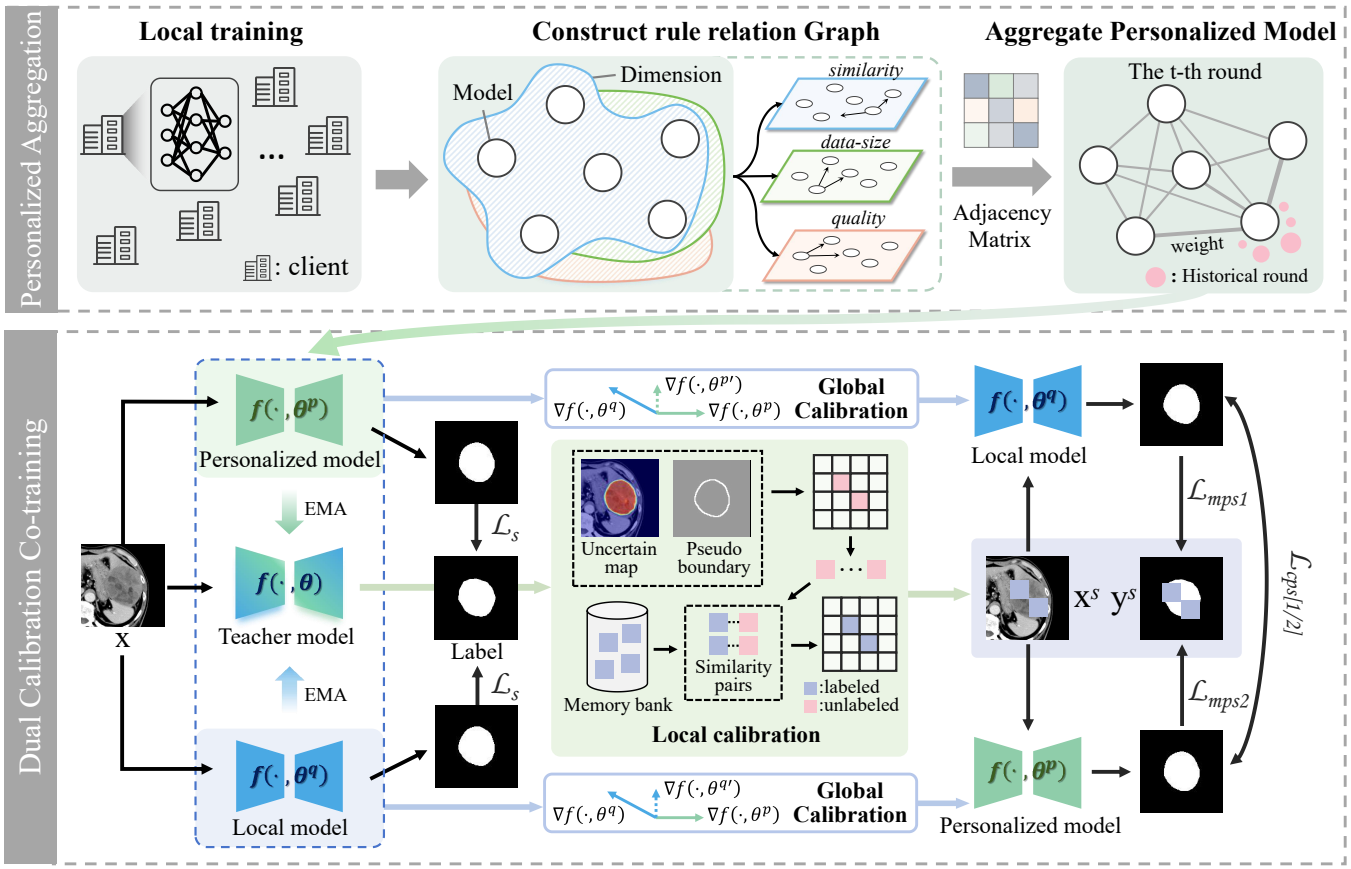


Figure 3: An overview of our proposed PFSSL framework. The upper side shows the federated learning architecture, where the server builds a rule relationship graph for multi-client models and aggregates a personalized model for each client. The lower side depicts the co-training process with dual calibration in semi-supervised learning, where local calibration corrects pseudo-labels to enhance the boundary discernment ability of the local model, and global calibration adjusts the gradient directions of the local model and the personalized model in real-time.

Federated Semi-Supervised Learning

Personalized Federated Semi-Supervised Learning (PFSSL) combines the advantages of personalized federated learning and semi-supervised learning to protect client data privacy while reducing annotation costs and addressing significant data heterogeneity issues. Recently, UM-pFSSL (Shi, Chen, and Zhang 2022) proposed client assistants to help unlabeled clients obtain trustworthy pseudo-labels. SemiPFL (Tashakori et al. 2023) considers the performance challenges of aggregating personalized models across multiple edge device users. However, these approaches are limited to specific scenarios and cannot be directly applied to medical image segmentation. In real-world medical scenarios, a single medical institution can annotate only a minimal amount of data. Our research is conducted in this context, where each client contains a small amount of labeled data and a large amount of unlabeled data.

Methodology

This section introduces our personalized federated semi-supervised learning framework, as depicted in Figure 3. First, we define the problem and introduce the aggregation process of the personalized model. Then, we describe in detail the optimization strategy of the dual calibration method with co-training.

Problem Settings

In our PFSSL setting, we assume there are K clients, where each client dataset contains a labeled dataset $D_k^l = \{(x_{k,i}^l, y_{k,i})|i = 1, ..., |D_k^l|\}$ and an unlabeled dataset $D_k^u = \{(x_{k,i}^u)|i = 1, ..., |D_k^u|\}$, usually $|D^l| \leq |D^u|$. The goal of PFSSL is to train a high-performance local segmentation model $f(\cdot, \theta)$ for each client.

Personalized Aggregation

We propose a federated learning personalized aggregation algorithm based on the rule relationship graph. Unlike traditional FL methods, the server receives the uploaded models and aggregates a unique personalized model for each client (Qi and Li 2024; Li et al. 2020). To determine the collaborative relationships among clients, we define a rule relationship graph $G(\mathcal{V}, \mathcal{E}, A)$, where the node set $\mathcal{V} =$

$\{c_1, c_2, ..., c_K\}$ representing all K clients, $\mathcal{E}$ is the set of edges representing relationships among the models of clients, and $\boldsymbol{A} \in \mathbb{R}^{K\times K}$ is a symmetric weighted adjacency matrix whose (i, j)th element quantifies the extent of collaboration between the ith and the jth clients.

Since the data distribution is inaccessible, we consider establishing a basic collaboration based on model similarity. Furthermore, we define two additional rules to jointly guide the collaboration among clients: the relative dataset size among clients and the model quality score. The model quality score is approximated by the segmentation prediction results of the client's personalized model. Then, the optimization objective for the i-th client on the server side is:

$$\begin{aligned} &\min_{\{\boldsymbol{A}_{ij}\}_j} \sum_j (\boldsymbol{A}_{ij} - (\rho g_j + \phi h_j))^2 - \alpha \sum_j \boldsymbol{A}_{ij} \cos(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \\ &\quad s.t. \sum_j \boldsymbol{A}_{ij} = 1, \forall i; \quad \boldsymbol{A}_{ij} \geq 0, \forall i, j; \quad \rho + \phi = 1 \end{aligned} \tag{1}$$

where $g_j = D_j / \sum_j^k D_j$ is the relative dataset size, h_j is the result calculated by model quality estimation function, $cos(\theta_i, \theta_j)$ is the cosine similarity between two models, α and ρ are the hyperparameters that influence the collaborative relationship based on model similarity and control the proportion of model rule, respectively. In the objective function, the first term accounts for the dataset size and model quality evaluation results, where more extensive or more accurate knowledge contributions are given higher collaboration strength. The second term considers model similarity to determine basic collaboration. The first constraint limits the overall collaboration budget of each client, and the second constraint requires all collaboration strengths to be positive. The third constraint defines the range of the rule.

Considering that personalized model aggregation is a dynamic process, we use the historical round information accumulated in the rule relationship graph to converge the current node weights (Zhu, Chen, and Yuan 2023). In each communication round t, we update the weight matrix A_{ij} by combining the current and past rule relationships and perform symmetric normalization. For the weight matrix A_{ij} :

$$A_{ij} = \mathcal{D}_{ii}^{-\frac{1}{2}} \cdot (\sum_{k=0}^{t} \mu^k \cdot A_{ij}^{t-k}) \cdot \mathcal{D}_{ii}^{-\frac{1}{2}} \tag{2}$$

where $\mathcal{D}_{ii} = diag(\sum_{k=0}^{t} \mu^k \cdot A_{ij}^{t-k})$ is the degree matrix, μ is the hyperparameter, and A^t represents the weight matrix at round t.

Dual-Calibrated Co-training

In this section, we present a dual calibration strategy within the co-training mechanism, which optimizes client models from local and global perspectives. We will further elaborate on the co-training mechanism and the calibration process in the following subsections.

Co-training Student-Teacher. In FSSL scenarios, relying on a single model may lead clients to prioritize local knowledge while neglecting global knowledge, resulting in over-personalization. Inspired by Mean Teacher (Tarvainen and Valpola 2017) and Cross Pseudo Supervision (Chen et al. 2021), we built a single-teacher-dual-student architecture with the local model $f(\cdot, \theta^q)$ and the personalized model $f(\cdot, \theta^p)$, which promotes the consistency of pseudo-labels by co-training with the dual students.

During the training process, gradient propagation is only performed within the student models, while the teacher model serves as a self-ensemble of the student model training results. Specifically, the teacher model is updated by the exponential moving average (EMA) of the balance weights from the student models: $\theta_i = \gamma\theta_{i-1} + (1-\gamma)(\beta\theta_i^q + (1-\beta)\theta_i^p)$, where i represents the local training iteration, γ is the EMA decay that controls the parameter update rate, and β is the balance coefficient of the student model ensemble. The β is calculated as follows:

$$\beta = \frac{1 - \mathcal{L}_{dc}(\theta_i^q; D^l)}{2 - \mathcal{L}_{dc}(\theta_i^q; D^l) - \mathcal{L}_{dc}(\theta_i^p; D^l)} \tag{3}$$

where $\mathcal{L}_{dc}$ represents the supervised dice loss (i.e., segmentation performance) of the two student networks on the labeled dataset. We choose the dice loss to control the balance coefficient, which determines the integration of more useful knowledge into the teacher model.

Local calibration. Differences in data distribution among clients may accumulate a large amount of heterogeneous pseudo-label noise in the SSL environment. We propose correcting pseudo-labels to calibrate the local model's ability to distinguish fuzzy boundaries (Pati et al. 2022), thereby improving the applicability of local models with SSL methods.

As shown in the local calibration module of Figure 3, the teacher model generates pseudo-labels and uncertainty maps using unlabeled data, where pseudo boundaries are obtained from the pseudo-labels by the boundary detection algorithm. We divide the uncertainty map into patches of size h × w in a block-wise manner and merge them with the pseudo boundary to form an uncertainty patch map that only contains the boundaries.

$$M_{i,j}^u = \sqrt{S(m_{i,j}^c, d = x)^2 + S(m_{i,j}^c, d = y)^2} \tag{4}$$

$$P_b^u = -\frac{1}{C}\frac{1}{h \times w} \sum_{b=1}^{h\times w} \sum_{c=1}^{C} (p_{i,j}^c \log(p_{i,j}^c) \cdot M_{i,j}^u) \tag{5}$$

where S is the Sobel boundary detection operator, d is the directional gradient, and $m_{i,j}^c$ is the classification result of the unlabeled image prediction $f(x_{i,j}^u, \theta)$ for pixel (i, j) on class c. In Eq(4), the pseudo boundary feature information $M_{i,j}^u$ is obtained through horizontal and vertical detection. In Eq(5), $p_{i,j}^c$ is the activation probability distribution of pixel (i, j) on the class c in the uncertainty map, and C is the total number of classes. Obviously, the b-th patch can reflect the model's confidence in the pixels within the patch.

Similarly, we calculate the boundary uncertainty patches for labeled data and select those with lower uncertainty as reliable knowledge to store in a fixed-capacity memory bank. Given the diversity of label scales and the differences in the quality and quantity of boundary patches, we dynamically

adjust the number of patch fusions to ensure the effective transfer of reliable knowledge.

$$h = min\{|P^l|, |P^u|\} \cdot \mathcal{P}_{score} \cdot \eta \tag{6}$$

where $|P^{[l/u]}|$ denotes the number of labeled or unlabeled boundary patches, we choose the minimum value of $|P^l|$ and $|P^u|$ as the replacement threshold, $\mathcal{P}_{score}$ represents the prediction performance of labeled data as the reliability of knowledge replacement, η is a hyperparameter used to control the number of replacements. We sort the pseudo-label boundary patches by uncertainty and select the top-h most uncertain (low-confidence) regions (red squares) to form similarity pairs with labeled data regions (blue squares) that are stored in the memory bank. Subsequently, the pseudo-labels are replaced based on the similar pairs, generating new images and pseudo-labels denoted as $x^s = Replace(x^l, x^u; P_b^l, P_b^u; h), y^s = Replace(y^l, y^u; P_b^l, P_b^u; h)$. The new samples are then used for supervised retraining.

The total loss $\mathcal{L}$ for the k-th client in local semi-supervised learning consists of two parts: supervised loss $\mathcal{L}_s$ and unsupervised loss $\mathcal{L}_u$, expressed as $\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_u$, where λ is a regularization parameter to balance the supervised and unsupervised learning losses. The supervised loss is formulated as,

$$\mathcal{L}_s = \frac{1}{|D_k^l|} \sum_{i=1}^{|D_k^l|} \left\{ \mathcal{L}_{dc}(f\left(x_i^l, \theta^q\right), y_i) + \mathcal{L}_{dc}(f\left(x_i^l, \theta^p\right), y_i) \right\} \tag{7}$$

Unsupervised loss includes the cross pseudo supervision and mean-teacher pseudo supervision, i.e., $\mathcal{L}_u = (\mathcal{L}_{cps1} + \mathcal{L}_{cps2}) + (\mathcal{L}_{mps1} + \mathcal{L}_{mps2})$, which are defined as follows:

$$\mathcal{L}_{cps[1/2]} = \frac{1}{|D_k^u|} \sum_{i=1}^{|D_k^u|} \mathcal{L}_{dc}(f(x_i^s, \theta^{[p/q]}), f(x_i^s, \theta^{[q/p]})) \tag{8}$$

$$\mathcal{L}_{mps[1/2]} = \frac{1}{|D_k^u|} \sum_{i=1}^{|D_k^u|} \mathcal{L}_{dc}(f(x_i^s, \theta^{[q/p]}), y_i^s) \tag{9}$$

where x_i^s and y_i^s are the fused images and pseudo-labels.

Global calibration. After each round of federated aggregation, the personalized model absorbs knowledge from multiple clients, which may lead to significant knowledge deviations from the local model. Furthermore, it potentially causes imbalances in the teacher model. We control the global direction by sensing the gradient directions of both the local model and personalized model in real time and guiding each other for calibration. Taking the local model as an example, the optimization for the i-th client is:

$$\min_{\boldsymbol{\theta}_i} f(\cdot, \boldsymbol{\theta}_i^q) = f(\cdot, \sum_j \boldsymbol{A}_{ij} \boldsymbol{\theta}_j^p) - \frac{\lambda}{2} \sum_j \boldsymbol{A}_{ij} \cos(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \tag{10}$$

The first term minimizes the loss of semi-supervised learning to pursue personalized model utility. The second term maximizes the cosine similarity between the local model and the personalized model to avoid excessive deviation and reduce the risk of overfitting.

Experiments

Experimental Setup

Datasets. We conducted comprehensive experiments on one private dataset and two public datasets, focusing on liver organs, skin cancer lesions, and colorectal polyps. The LiverSeg23 private dataset, sourced from six hospitals in China, contains 1,139 patient samples, 11,701 CT images, and corresponding annotations. Since the ISIC-2018 (Codella et al. 2019) dataset for the skin cancer lesion task follows independent and identically distributed (IID), we use Dirichlet distribution to generate non-IID data partitions of the 3594 image files among four clients. The polyp segmentation task uses colonoscopy images from five different data sources for the experiment (Jha et al. 2020; Bernal et al. 2015). In our studies, each sub-dataset was treated as an individual client's private dataset and randomly split into 80% for training and 20% for testing. Finally, we merged all test sets into a global test set to validate the performance of the aggregated model.

Implementation Details. We implemented our framework using the Pytorch library on Linux system. We employed U-Net (Ronneberger, Fischer, and Brox 2015) as the base model for each client, with a pre-trained ResNet-34 (He et al. 2016) as the backbone network. All training images were resized to 256 $\times$ 256. Each local model is trained via an AdamW optimizer with a batch size of 8, Adam momentums of 0.9 and 0.999, and a fixed learning rate uniformly as 1e-4. We set the federated hyperparameters ρ = 0.8, and μ = 0.1. In the semi-supervised setting, the labeled data ratio factor is set to 0.3, the patch size is set to 8 and the patch replacement number hyperparameter η is set to 1. We trained 20 federated rounds in total or until the model has converged stably, where the local epoch is set as 10 by default. Additionally, the basic settings for all comparison methods were kept consistent with our method.

Evaluation Metrics. We employ four widely used metrics to evaluate segmentation performance, including dice coefficient (DC), jaccard index (JC), hausdorff distance (HD95), and average symmetric surface distance (ASSD).

Comparisons with State-of-the-arts

Comparison Methods. We compare the following methods in experiments. FedIRM (Liu et al. 2021) and HSSF (Ma et al. 2024b) are state-of-the-art FSSL methods. FedAvg+MT (McMahan et al. 2017; Tarvainen and Valpola 2017), PgFed+dhc (Luo et al. 2023; Wang and Li 2023), FedProx+FixMatch (Li et al. 2020; Sohn et al. 2020), and FedU+bcp (Dinh et al. 2022; Bai et al. 2023) are naive combinations of FL or PFL methods and SSL methods. We also compare our method against FedAvg combined with supervised methods, using FedAvg with fully labeled data and 10% labeled data as the upper and lower bounds, respectively (Luo et al. 2021b; Luo 2020).

Quantitative Comparisons. Table 1 reports the quantitative comparison results of our method and other state-of-the-art methods on the medical datasets. For liver segmentation results, we can observe that our method achieves 84.16%, 85%, 7.95, and 3.14 on average in the four metrics, which outperforms the latest competitor HSSF with clear

Methods	LiverSeg23				ISIC2018				Polyp			
	DC ↑	JC ↑	HD95 ↓	ASSD ↓	DC ↑	JC ↑	HD95 ↓	ASSD ↓	DC ↑	JC ↑	HD95 ↓	ASSD ↓
HSSF	71.47	75.45	12.10	4.75	84.76	83.36	10.25	3.37	41.97	62.88	22.34	11.32
FedIRM	82.74	83.83	8.67	3.34	86.77	84.85	9.98	3.39	60.83	71.82	18.15	7.56
PgFed+dhc	81.64	82.98	8.65	3.51	86.22	84.38	10.66	3.68	66.75	74.55	21.14	8.17
FedU+bcp	81.13	82.42	10.45	3.61	85.71	83.53	11.64	3.79	57.75	69.47	23.26	10.21
FedAvg+mt	77.39	79.88	9.54	3.78	86.35	84.39	11.22	3.67	59.68	72.33	25.54	11.08
FedProx+FixMatch	71.00	75.87	10.96	4.63	87.68	85.74	10.36	3.32	63.11	73.33	22.74	9.79
Fedavg(upper-bound)	88.27	88.33	6.92	2.45	89.16	87.30	9.69	2.97	71.72	77.51	19.83	7.81
Fedavg(lower-bound)	49.01	63.22	9.54	4.85	81.61	80.34	15.61	5.10	40.03	58.62	30.69	12.98
ours	**84.16**	**85.00**	**7.95**	**3.14**	**89.44**	**87.58**	**8.70**	**2.84**	**73.86**	**79.79**	**17.71**	**7.51**

Table 1: The average results of LiverSeg23, ISIC2018, and Polyp datasets under multi-client heterogeneous data partition. The results reported that our method performs relatively better than all methods.

LC	PA	GC	DC ↑	JC ↑	HD95 ↓	ASSD ↓
×	×	×	77.39	79.88	9.54	3.78
✓	×	×	80.76	81.64	8.94	3.65
×	✓	×	81.77	83.07	8.76	3.38
×	×	✓	77.81	80.28	9.39	3.75
✓	✓	×	83.48	84.51	8.18	3.19
✓	×	✓	81.61	82.90	8.89	3.54
×	✓	✓	74.15	77.78	11.24	4.44
✓	✓	✓	**84.16**	**85.00**	**7.95**	**3.14**

Table 2: Ablation study on the effectiveness of each component using the LiverSeg23 dataset.

Methods	DC ↑	JC ↑	HD95 ↓	ASSD ↓
baseline	81.21	82.68	8.60	3.41
MiDSS	83.73(+2.52)	84.61(+1.93)	8.19	3.06
UCMT	85.08(+3.87)	85.60(+2.92)	8.07	2.98
Ours	**87.20(+5.99)**	**87.41(+4.73)**	**7.40**	**2.62**

Table 3: Ablation study on the effectiveness of local calibration using the LiverSeg23 dataset.

improvements of 12.69% in DC and 9.55% in JC. Moreover, simply combining personalized federated learning with semi-supervised learning does not yield better results, even with state-of-the-art methods. The main reason is that differences in individual data distributions cannot be sufficiently adapted in semi-supervised learning, thereby hindering the optimization of the aggregated model. In contrast, our PFSSL method effectively combines federated learning and semi-supervised learning and utilizes dual calibration to optimize the model's learning ability.

For skin lesion and polyp segmentation results, we can see that our method outperforms both state-of-the-art and classical methods in most evaluation metrics, even surpassing the upper bound. Especially in the polyp segmentation task, the significant distribution differences in data sources cause most methods to suffer from severe data drift issues, performing even close to the lower performance bound. This issue arises because forcibly aggregating significantly different client models misleads the learning of the local model, resulting in a series of performance issues. Instead, our method achieves consistent performance gains regardless of the segmentation tasks. To more intuitively demonstrate the superiority of our method, Figure 4 visualizes some example results of liver segmentation, skin lesion segmentation, and polyp segmentation.

Ablation Studies

Effectiveness of Components. We conducted a quantitative comparison to evaluate the effectiveness of each component in our method on the liver organ segmentation task. We applied MT and FedAvg in clients to form a fundamental FSSL method as the baseline. Similarly, we substituted the comparison methods that do not use our components with either MT or FedAvg. As shown in Table 2, the baseline method yielded poorer segmentation performance, which may be due to the negative effects of client models with noisy data on the convergence of the global model. Introducing Local Calibration (LC), Personalized Aggregation (PA), or Global Calibration (GC) individually led to significant improvements over the baseline, demonstrating the effectiveness of each component we proposed. We further combined the components sequentially to verify the effectiveness of these combinations. The experimental results clearly show that applying PA or GC based on LC achieves consistent improvements. However, the performance of GC based on PA decreased, which we suspect is due to the instability of local models affecting the global calibration. Finally, we achieved final results with 84.16% in Dice and 85% in HD95 by seamlessly integrating each proposed component to obtain our proposed method, significantly outperforming other methods on the liver segmentation task.

Effectiveness of Local Calibration. Considering the proposed dynamic boundary patch fusion strategy of the local calibration method is also a data fusion method for semi-supervised learning. We compared it with the state-of-the-art fusion methods, MiDSS (Ma et al. 2024a) and UCMT (Shen et al. 2023), to verify the reliability of the calibrated pseudo-labels. In the semi-supervised scenario, we train a supervised model using labeled data as the baseline, then select a certain amount of data samples as the un-fusion dataset. Next, we employed these methods to generate the fused data, which was then used to retrain the supervised model. As shown in Table 3, our method improved by 5.99% and

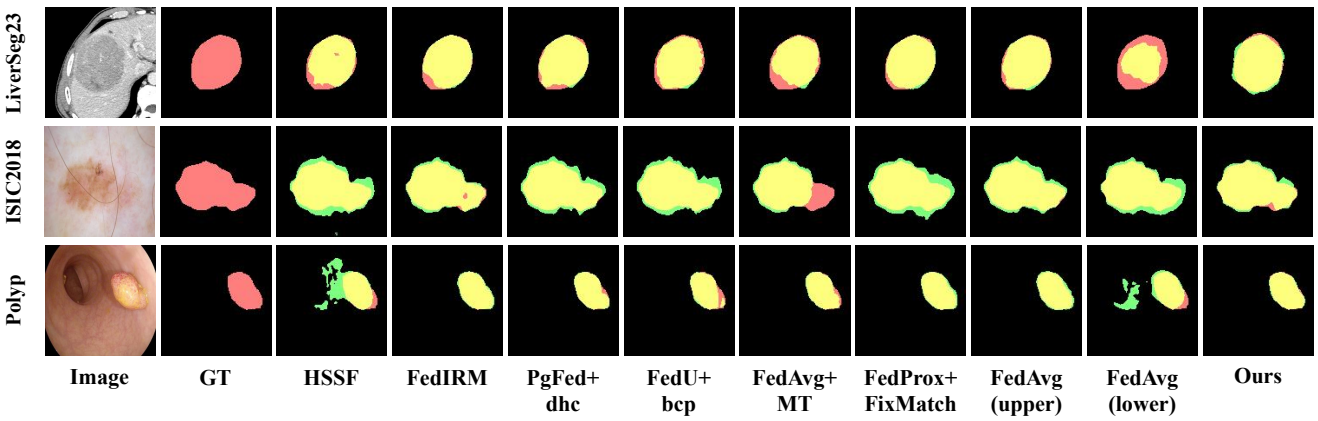
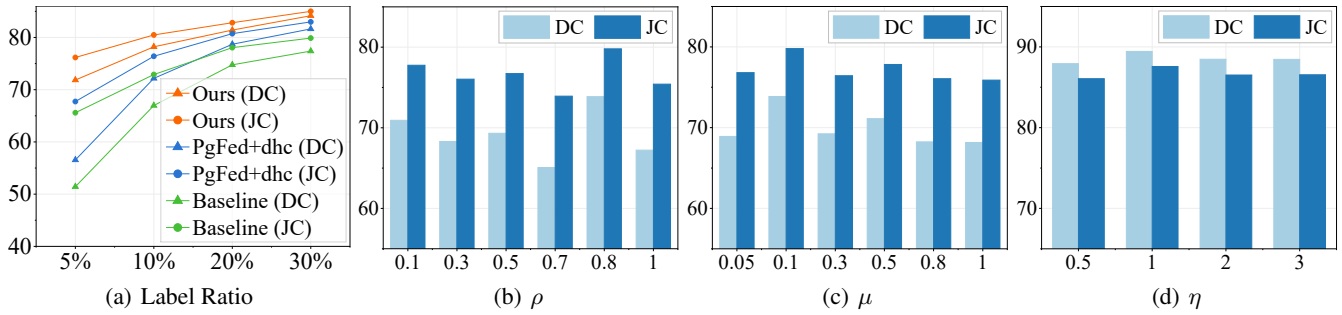


Figure 4: Visual comparison with different state-of-the-art methods on the LiverSeg23, ISIC2018, and Polyp datasets. The red, green, and yellow pixels indicate the ground truths, predictions, and their overlapping regions, respectively.



(a) Label Ratio (b) ρ (c) μ (d) η

Figure 5: Analysis of the impacts of labeled data ratio and different hyper-parameters. (a) Effect of the labeled data ratio on the LiverSeg23 dataset and using FedAvg+MT as our baseline method. (b) and (c) illustrate the effect of different ρ and μ values on the Polyp dataset using our method. (d) Effect of different η values on the ISIC2018 dataset using our method. We have utilized the DC and JC as the two main metrics for comparative analysis.

4.73% in DC and JC metrics compared with the baseline, significantly surpassing the improvements of other methods. This result indicates that the pseudo-labels generated by our method are effective and can convey valuable knowledge for model retraining.

Labeled Data Ratio. We evaluated our method on the LiverSeg23 dataset. In addition to adjusting the ratio of labeled data, we also compared our method with the best-performing method and the baseline method. We replicated the labeled data to match the quantity of unlabeled data, ensuring a balanced data volume. As shown in Figure 5(a), our method outperforms the PgFed+dhc and baseline methods. Moreover, both DC and JC metrics show an increasing trend as the ratio of data increases, which indicates that our method is not affected by the ratio of labeled data in local clients.

Hyper-parameters. Note that our method has three main hyperparameters: the personalized aggregation rule parameter ρ, the historical information coefficient μ, and the number of pseudo-label boundary patch fusion η. We first conducted experiments on the polyp dataset with parameters ρ and μ. The experimental results shown in Figure 5(b) and 5(c), which indicate that set $\rho = 0.8$ and $\mu = 0.1$ is optimal. Furthermore, we conducted experiments on the ISIC dataset to determine the optimal boundary patch fusion number η. As shown in Figure 5(d), continuously increasing the value of η does not improve overall segmentation performance, as replacing too many boundary patches may convey incorrect knowledge. Therefore, we chose to set the default value of η to 1.

Conclusion

In this paper, we address two thorny challenges in FSSL, the data heterogeneity among clients and the adaptability of the semi-supervised method. We ingeniously introduce personalized federated learning to construct rule relationship graphs across multiple clients to customize a personalized model for each client and co-training with the private local model. Additionally, we designed a dual calibration strategy to further calibrate the local and personalized models. We evaluated our method on both private and public medical datasets, and the results demonstrated the effectiveness and superiority of our approach.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grants (W2411054, U21A20521 and 62271178), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23_2524), National Foreign Expert Project of China under Grant (G2023144009L), Zhejiang Provincial Natural Science Foundation of China (LR23F010002), Wuxi Health Commission Precision Medicine Project (J202106), Jiangsu Provincial Six Talent Peaks Project (YY-124), and the construction project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400).

References

Antunes, R. S.; André da Costa, C.; Küderle, A.; Yari, I. A.; and Eskofier, B. 2022. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4): 1–23.

Bai, Y.; Chen, D.; Li, Q.; Shen, W.; and Wang, Y. 2023. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11514–11524.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43: 99–111.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2613–2622.

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*.

Dinh, C. T.; Vu, T. T.; Tran, N. H.; Dao, M. N.; and Zhang, H. 2022. A new look and convergence rate of federated multitask learning with laplacian regularization. *IEEE Transactions on Neural Networks and Learning Systems*.

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.

Hung, W.-C.; Tsai, Y.-H.; Liou, Y.-T.; Lin, Y.-Y.; and Yang, M.-H. 2018. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*.

Jeong, W.; Yoon, J.; Yang, E.; and Hwang, S. J. 2020. Federated semi-supervised learning with inter-client consistency & disjoint learning. *arXiv preprint arXiv:2006.12097*.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; De Lange, T.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, 451–462. Springer.

Kalra, S.; Wen, J.; Cresswell, J. C.; Volkovs, M.; and Tizhoosh, H. R. 2023. Decentralized federated learning through proxy model sharing. *Nature communications*, 14(1): 2899.

Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Kulkarni, V.; Kulkarni, M.; and Pant, A. 2020. Survey of personalization techniques for federated learning. In *2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4)*, 794–797. IEEE.

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. Atlanta.

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.

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, Q.; Dou, Q.; Yu, L.; and Heng, P. A. 2020. MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. *IEEE transactions on medical imaging*, 39(9): 2713–2724.

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.

Liu, Y.; Wu, H.; and Qin, J. 2024. FedCD: Federated Semi-Supervised Learning with Class Awareness Balance via Dual Teachers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3837–3845.

Luo, J.; Mendieta, M.; Chen, C.; and Wu, S. 2023. Pgfed: Personalize each client's global objective for federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3946–3956.

Luo, M.; Chen, F.; Hu, D.; Zhang, Y.; Liang, J.; and Feng, J. 2021a. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34: 5972–5984.

Luo, X. 2020. SSL4MIS. https://github.com/HiLab-git/SSL4MIS.

Luo, X.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Chen, N.; Wang, G.; and Zhang, S. 2021b. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 318–329. Springer.

Ma, Q.; Zhang, J.; Qi, L.; Yu, Q.; Shi, Y.; and Gao, Y. 2024a. Constructing and Exploring Intermediate Domains in Mixed Domain Semi-supervised Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11642–11651.

Ma, Y.; Wang, J.; Yang, J.; and Wang, L. 2024b. Model-Heterogeneous Semi-Supervised Federated Learning for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*.

Marfoq, O.; Neglia, G.; Bellet, A.; Kameni, L.; and Vidal, R. 2021. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34: 15434–15447.

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.

Pati, S.; Baid, U.; Edwards, B.; Sheller, M.; Wang, S.-H.; Reina, G. A.; Foley, P.; Gruzdev, A.; Karkada, D.; Davatzikos, C.; et al. 2022. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1): 7346.

Qi, F.; and Li, S. 2024. Adaptive Hyper-graph Aggregation for Modality-Agnostic Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12312–12321.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Shen, Z.; Cao, P.; Yang, H.; Liu, X.; Yang, J.; and Zaiane, O. R. 2023. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. *arXiv preprint arXiv:2301.04465*.

Shi, Y.; Chen, S.; and Zhang, H. 2022. Uncertainty minimization for personalized federated semi-supervised learning. *IEEE Transactions on Network Science and Engineering*, 10(2): 1060–1073.

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.

Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2022. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12): 9587–9603.

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.

Tashakori, A.; Zhang, W.; Wang, Z. J.; and Servati, P. 2023. SemiPFL: Personalized semi-supervised federated learning framework for edge intelligence. *IEEE Internet of Things Journal*, 10(10): 9161–9176.

Wang, H.; and Li, X. 2023. Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 582–591. Springer.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International conference on machine learning*, 7164–7173. PMLR.

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

Zhu, M.; Chen, Z.; and Yuan, Y. 2023. FedDM: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting. *IEEE Transactions on Medical Imaging*, 42(6): 1632–1643.