Federated Semi-Supervised Learning for Medical Image Segmentation via Pseudo-Label Denoising

Liang Qiu, Jierong Cheng, Huxin Gao, Wei Xiong, and Hongliang Ren

***Abstract*—Distributed big data and digital healthcare technologies have great potential to promote medical services, but challenges arise when it comes to learning predictive model from diverse and complex e-health datasets. Federated Learning (FL), as a collaborative machine learning technique, aims to address the challenges by learning a joint predictive model across multi-site clients, especially for distributed medical institutions or hospitals. However, most existing FL methods assume that clients possess fully labeled data for training, which is often not the case in e-health datasets due to high labeling costs or expertise requirement. Therefore, this work proposes a novel and feasible approach to learn a Federated Semi-Supervised Learning (FSSL) model from distributed medical image domains, where a federated pseudo-labeling strategy for unlabeled clients is developed based on the embedded knowledge learned from labeled clients. This greatly mitigates the annotation deficiency at unlabeled clients and leads to a cost-effective and efficient medical image analysis tool. We demonstrated the effectiveness of our method by achieving significant improvements compared to the state-of-the-art in both fundus image and prostate MRI segmentation tasks, resulting in the highest Dice scores of 89.23% and 91.95% respectively even with only a few labeled clients participating in model training. This reveals the superiority of our method for practical deployment, ultimately facilitating the wider use of FL in healthcare and leading to better patient outcomes.**

***Index Terms*—Federated learning, federated pseudo-labeling strategy, medical image segmentation.**

Manuscript received 1 July 2022; revised 28 November 2022, 13 March 2023, and 11 April 2023; accepted 2 May 2023. Date of publication 8 May 2023; date of current version 5 October 2023. This work was supported in part by Hong Kong Research Grants Council Collaborative Research Fund under Grants CRF C4026-21GF and CRF C4063-18G, in part by the General Research Fund under Grants GRF 14211420 and GRF 14216022, in part by the Shun Hing Institute of Advanced Engineering under Grant BME-p1-21/8115064, and in part by the CUHK; and Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) under Grant 202108233000303. *(Corresponding authors: Hongliang Ren; Wei Xiong.)*

Liang Qiu is with the Department of Radiation Oncology, Stanford University, Stanford, CA 94305 USA (e-mail: qiuliang@stanford.edu).

Jierong Cheng and Wei Xiong are with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632 (e-mail: chengjr@i2r.a-star.edu.sg; wxiong@i2r.a-star.edu.sg).

Huxin Gao is with the Department of Biomedical Engineering, National University of Singapore, Singapore 119077 (e-mail: e0343967@u.nus.edu).

Hongliang Ren is with the Department of Electronic Engineering, Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong, Hong Kong. (e-mail: hlren@ieee.org).

Digital Object Identifier 10.1109/JBHI.2023.3274498

I. Introduction

TODAY'S healthcare has been undergoing a digital revolution, penetrating almost every aspect of our daily lives. The concept of digital health or e-Health refers to the use of information technology equipped with cutting-edge resources to provide more efficient management and easier accessibility to patients, and of course, reliable and affordable treatment [1], [2]. Versatile technologies such as tele-health, wearable devices, augmented reality and digital record constitute a powerful e-Health ecosystem, leading to a series of benefits, e.g., medical monitoring improvement, easier decision-making for clinician, and more informed patients. During the ongoing coronavirus disease 2019 (COVID-19) crisis, there are greater demand for digital health advancements, along with potential ethical or technical challenges [3], [4]. Particularly, medical diagnosis and treatment have been effectively promoted by artificial intelligence based on big data in healthcare to be widely effective [5], [6], [7], [8], [9]. As the continuous improvement of electronic health record (EHR) and electronic medical record (EMR) systems, massive e-Health digital data can be accessed or shared easily under certain privacy protection policies, which provides a convenient platform for AI-based healthcare. However, these learning-based models may overfit subtle institutional data biases and generalize poorly to other institutional data due to widespread heterogeneous data distribution. A natural way to improve the model generalizability is through collaborative learning, where data from multiple institutions with great diversity are leveraged to train a single powerful model. With this prediction tool, instantaneous e-healthcare services can be realized or enhanced to directly help patient care. Currently, collaborative data sharing (CDS) is the common paradigm for multi-institution collaboration, which requires institutions to share their patient data to a central location for model training [10]. However, data privacy, technical and ethical constraints, and data ownership concerns make it intractable for a large number of institutions to collaborate in practice, especially in an international setting, necessitating the search for alternative approaches.

Recently FL as a decentralized machine learning paradigm provides a promising privacy-preserving solution to collaboratively learn a generalized prediction model across multi-site data instead of over centralized information, effectively mitigating data scarcity and distribution bias [11], [12], [13], [14]. It learns a global consensus model on the central server by aggregating the contributions from each local client, i.e., averaging parameters, and shares the updated model to all the participants for further parallel local training in a circulating process until the model reaches the desired accuracy. Each iteration of this circulating process including parallel training, aggregation update and new parameter distribution is known as a federated round. Given the significance of privacy protection, the benefits of FL filter out into the wider healthcare ecosystem [15], [16], [17]. It opens up the possibility for different hospitals, healthcare institutions and medical research centers to collaborate on building a robust model based on distributed datasets representing a wider demographic of patients, yet with no leakage of sensitive private patient information. This finding has the potential to shift the paradigm of multi-institutional collaborations, and model training using federated learning across multiple datasets yields comparable results to model training using CDS. Although FL revolutionizes the training paradigm of artificial intelligence and overcomes the multi-site data access restriction due to strict regulatory policy on data privacy, local clients without expert annotations still cannot join and contribute to the FL process effectively because of the supervised setting in most existing paradigms. It is a common limitation in healthcare due to high annotation costs and adequate expertise requirements that ordinary medical units probably cannot afford, especially for segmentation tasks needing pixel-level labeling, which heavily discourages the widespread adoption of FL [18], [19]. Therefore, conventional FL methods are inapplicable in those realistic medical scenarios, leading to a practical FL problem involving distributed unlabeled datasets, namely federated semi-supervised learning (FSSL), aiming to utilize extensive unlabeled clients to further facilitate FL.

Considering the generally minor domain shifts in distributed datasets instead of the cross-modality discrepancy, we can exploit the well-trained source model on labeled clients as an important supervision base for model self-training on the target domains [20]. Therefore, tackling label noise and producing reliable pseudo labels for unlabeled clients is a crucial step in FSSL scenarios. Then the FL model could benefit from local parameter aggregation supervised by both labeled and pseudo-labeled data. In this article, we conducted the first rigorous investigation of pseudo-label denoising strategy in federated paradigm and proposed a novel and complete FSSL framework for medical image segmentation across distributed big data, which can effectively alleviate privacy concerns and fully utilize all available data to enable instantaneous e-healthcare services.

Our main contributions are highlighted as follows:

1) We propose a novel FSSL method for multi-site medical image segmentation tasks, which leverages a federated pseudo-labeling strategy with uncertainty and prototype estimations to predict reliable pseudo labels, effectively facilitating model learning from unlabeled clients.
2) Extensive experiments have been implemented in the retinal fundus image segmentation and prostate MRI segmentation tasks with cross-site domain shifts, validating the superior performance of our method against the state-of-the-art.
3) The effect of participating labeled client number has been further investigated, which demonstrates the superiority of our method in segmentation tasks even with a small number of labeled clients, facilitating the wider use and practical deployment in instantaneous healthcare services.

II. Related Work

A. FL in Medical Imaging

To facilitate large-scale multi-institutional cooperations, FL as a promising data-private collaborative learning approach has been explored in several studies, when facing massive e-health digital data from EHR or HER systems. The first use of FL for medical imaging was introduced for multi-site brain tumor segmentation, which demonstrated its effectiveness with comparable performance against CDS methods [21]. To further investigate the potential data leakage and data imbalance issues, differential-privacy techniques and weight sharing strategies were applied to enhance the model capability [22]. In addition, to comprehensively address the substantial bandwidth consumption and data leakage problem, a novel ensemble attention distillation algorithm was developed, which can efficiently gain knowledge from various decentralized data sources and provide superior prediction performance [23]. Moreover, a large and real-world healthcare study based on the data from 20 institutes across four continents was conducted to develop an FL model, called EXAM, for clinical prediction of patients with COVID-19, which was validated to be more accurate and robust compared with locally trained models. It practically demonstrated the feasibility of FL for rapid multi-site collaboration without sharing data between institutions [24]. Likewise, another large-scale FL study was performed among 10 institutions for brain tumor segmentation, which not only revealed its powerful prediction performance within the original institution group, also showed the generalizability of FL models outside federation [25]. Although larger and more diverse datasets play a significant role in improving the generalizability of FL approaches, it could still be suboptimal due to heterogeneous data distribution with domain shifts. To relieve the problem, a recent literature has proposed two domain adaptation methods in FL architectures for multi-site fMRI analysis, namely mixture of experts and adversarial domain alignment [26]. Similarly, an FL model for MRI reconstruction was proposed to circumvent this challenge by aligning data distribution in the latent feature space between the source domain and the target domain [27]. However, these methods typically require the target prior information for model adaptation, not suitable for unseen data sources outside the federation. Instead, a federated domain generalization method called FedDG was proposed, which can exchange the distribution information across clients within the frequency space to learn generalizable parameters [28]. However, the aforementioned FL

methods are all in supervised settings and may be infeasible when lacking enough expert annotations.

B. FSSL in Medical Imaging

To address the annotation deficiency problem in computer vision community, semi-supervised learning (SSL) is a fundamental solution by leveraging the supervision from both the labeled data and the potentially valuable information embedded in the unlabeled data [29]. Likewise, SSL has also been widely used in medical image analysis due to today's big data healthcare [30], [31], [32]. Thus, combining the existing SSL methods with the FL paradigm could facilitate the design of FSSL models. For example, an FSSL approach called Federated Matching was proposed to fully exploit the large distributed data without any accompanying labels, where the inter-client consistency loss was designed to regularize the local models with helper agents selected from a global server based on model similarity [33]. In addition, contrastive representation learning was integrated with FL to facilitate semi-supervised learning particularly in medical imaging scenarios. It first pre-trains the model on a large amount of unlabeled data to learn transferable representations, then fine-tunes using limited annotations for downstream tasks [34], [35]. This strategy assumes that both labeled and unlabeled data exist in each client while ignoring the totally unlabeled clients we are concerned about. Moreover, an FSSL method for COVID region segmentation was proposed using disjoint learning with parameter decomposition [36]. It exploited labeled clients to supervise the corresponding local training process and imposed perturbation-invariant regulation on unlabeled clients with pseudo-labeling technique, then gradients updated from both aspects were utilized for model aggregation. In its experimental setting, two of three datasets are utilized for FSSL and the other is exploited to evaluate the model's generalizability, which may be insufficient to support the FSSL scenario, where more unlabeled clients should be included in SSL. Notably, those proposed methods achieved moderate improvements over their counterparts trained with labeled images only. This is because consistency regulation is heavily dependent on reliable predictions for unlabeled client data, while noisy pseudo labels are unavoidable because of possible domain shifts in federated scenarios, which may directly affect the quality of the resulting models. Recently, an FSSL method for medical image classification has been developed based on inter-client relation matching, which exploits the inherent disease relationships independent of different hospitals [37]. At unlabeled clients, pseudo labels are generated from model predictions mainly for estimating the disease relation matrix. By minimizing the inter-client relation matching loss constructed by Kullback–Leibler (KL) divergence of the relation matrixes between labeled and unlabeled clients, the knowledge from labeled clients can be exploited indirectly to facilitate learning at unlabeled clients. Validation has been performed on intracranial hemorrhage and skin lesion classification tasks, but the data of each local client are randomly partitioned from one large dataset to simulate the FL setting, which may not practically represent the cross-client domain shifts. In addition, the direct transferability of this method to more complicated applications, e.g., medical image segmentation, seems intractable. In general, it is still challenging and complex to fully exploit unlabeled medical data to reduce the annotation burden under various distributed learning settings or scenarios. Particularly, we could find there are very limited existing FSSL methods for multi-site medical image segmentation, and the data distribution shift between labeled and unlabeled clients under the FL setting is still an unsolved problem. Therefore, we introduced a federated pseudo-labeling strategy to further improve the pseudo-label reliability for unlabeled data, and validated its significant effectiveness based on two medical segmentation tasks with much more participating local clients from different sources with obvious domain shifts against [36] to support the FSSL setting, where the effect of different unlabeled client number is particularly investigated to indicate the superior capability of our method.

III. Method

In our FSSL setting, a set of m label clients is given with corresponding datasets $D_L = \{D^1, D^2, \ldots, D^m\}$, where each client C_l contains N_l image/label pairs denoted as $D^l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$, and n unlabeled clients with corresponding datasets $D_U = \{D^{m+1}, D^{m+2}, \ldots, D^{m+n}\}$ where each client C_u contains N_u images denoted as $D^u = \{(x_i^u)\}_{i=1}^{N_u}$ (usually $|D_L| \leq |D_U|$). The medical data from each client are invisible to the central server and the other clients due to data privacy policy. The aim of our FSSL method is to learn a generalized global federated model f_θ by jointly exploiting the knowledge from both labeled and unlabeled clients with potential domain shifts, as shown in Fig. 1. Instead of aggregating the parameters simultaneously from all the clients, we divide our FSSL scheme into federated supervised learning (FSL) and federated denoised self-training (FDST), both of which employ FegAvg [38] as federated learning backbone. Specifically, FSL model f_ξ aggregates the local supervised model parameters from labeled clients and provides reliable pseudo labels denoised by uncertainty and prototype estimations to train the FDST model f_ψ for unlabeled clients. The global FSSL model will aggregate updates from both sides and broadcast the parameters to each client. The training procedure alternates between updating the FSSL model and refining the generated pseudo labels. The detailed algorithm flow is shown in Algorithm 1.

A. FSL at Labeled Clients

Given the expert annotation knowledge existing at labeled clients, we can train our FSL model following the standard FL paradigm involving the communication between the FSL model on a central server and the m labeled clients. We adopt the widely-used federated averaging algorithm FegAvg [38] as the FSL model backbone. In our scenario, each labeled client C_l first performs supervised learning on dataset D^l independently to obtain the trained model f_{ξ_l}, by minimizing the local cross-entropy loss per synchronization round, shown as follow:

$$\mathcal{L}_L(\xi_l) = CrossEntropy(y_i^l, f_{\xi_l}(x_i^l)) \tag{1}$$

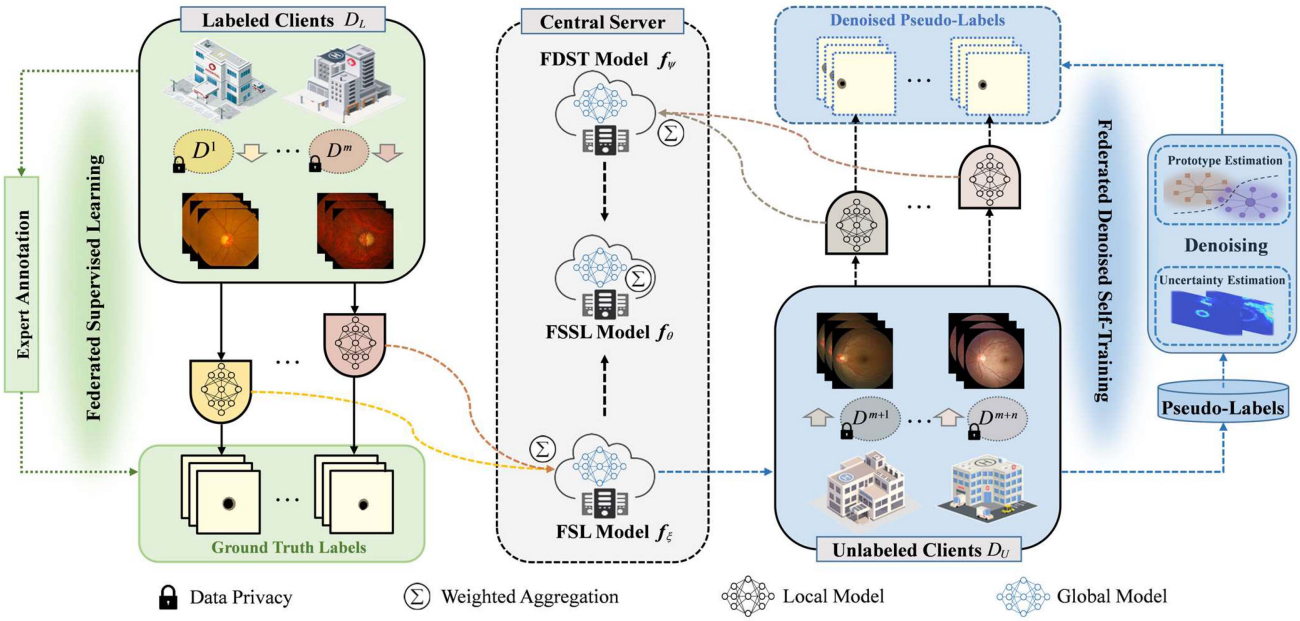

Fig. 1. Illustration of our FSSL framework at both labeled and unlabeled clients. Our FSSL scheme is composed of FSL and FDST, where FSL model f_ξ aggregates the local model parameters from labeled clients and provides denoised pseudo labels to train the FDST model f_ψ for unlabeled clients. The global FSSL model will aggregate updates from both sides and broadcast the parameters to each local client. The training procedure alternates between updating the FSSL model and refining the generated pseudo labels.

where the parameter ξ_l in each local model f_{ξ_l} is the optimization target. Then the server conducts the aggregation to get the FSL model parameters ξ through weighted summation in proportional to the size of each labeled dataset D^l, namely $\xi = \sum_{l=1}^{m} (N_l/N) \times \xi_l$, where $N = \sum_{l=1}^{m} N_l$. Subsequently, the new model parameters ξ will be sent back to all the labeled clients to update their local models. This process repeats until the FSL model converges. Through several rounds of server-client communication, the global model could be boosted with better generalizability, and the performance of the local models would also be further enhanced.

B. FDST at Unlabeled Clients

In practical situations, many unlabeled clients are impossible to contribute to the traditional FL paradigm to learn a generalization model. Despite lack of annotations, unlabeled clients with enormous embedded information should also be fully exploited in conjunction with labeled clients to learn superior segmentation models. Although consistency regularization on unlabeled clients could effectively assist FSSL [36], the unreliable pseudo annotations may hinder the collaborative learning effect due to potential domain shifts, even lead to model learning failure. Therefore, we innovatively introduce a pseudo-labeling strategy into this much more complex federated paradigm to combat the issue. After obtaining the denoised pseudo labels, we can utilize those pseudo labels to supervise local model training at unlabeled clients and perform FL following the similar procedure as described in FSL. The technical details will be elaborated as follows.

1) Plain Federated Pseudo-Labeling Generation: Given potential domain shifts across different clients, the FSL model learned from multiple labeled clients will possess stronger generalizability and provide more meaningful pseudo labels for unlabeled clients. Specifically, each unlabeled client first retrieves the well-trained FSL model f_ξ with weights ξ from the central server, then creates the pseudo label $\hat{y}_i^u$ based on the prediction probability $p_i^u = f_\xi(x_i^u)$ of each corresponding unlabeled image $x_i^u \in D^u$ for C-class segmentation, shown as follow:

$$\hat{y}_i^u = I_A\left(p_i^u > \lambda\right) \tag{2}$$

where I_A is the indicator function with the probability threshold parameter λ to generate the pseudo label $\hat{y}_i^u$.

2) Federated Pseudo-Label Denoising Strategy: The pseudo labels generated for unlabeled clients are unavoidably noisy because of domain biases, even predicted by the FSL model with better robustness and generalizability using (2). Unreliable pseudo labels may lead to inaccurate local models for those unlabeled clients and even disturb the parameter aggregation of the global model due to accumulated local model errors. Considering tremendous data collected from HER or EMR system for federated learning, manual participation for pseudo-label selection or refinement using active learning may be inappropriate because of limited effort and time. Therefore, denoising pseudo labels for high-quality labels is the key step to guarantee the effectiveness and efficiency of our FSSL model. Although pseudo-label refinement has been studied in in SSL, its application in federated paradigm with more challenging data heterogeneity has not been investigated. Motivated by the

Algorithm 1: FSSL for Medical Image Segmentation via Pseudo-Label Denoising.

Input: Labeled private datasets $D_L = \{D^1, D^2, \ldots, D^m\}$ where $D^l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ corresponding to C_l, unlabeled private datasets $D_U = \{D^{m+1}, D^{m+2}, \ldots, D^{m+n}\}$ where $D^u = \{(x_i^u)\}_{i=1}^{N_u}$ corresponding to C_u, global federated model f_θ, FSL model f_ξ, FDST model f_ψ.

Output: optimal ξ_l and ψ_u for each client C_l and C_u.

Initialize global model parameters $\theta \leftarrow \theta_0$
Send θ to labeled local clients for initialization $\xi_l^0 \leftarrow \theta$
for each round $r = 1, 2, \ldots, R$ **do** // FSL at LabeledClients
  for each client C_l **in parallel do**
    $\xi_l^r \leftarrow$ **ClientUpdate**(D^l, ξ_l^{r-1})
  end for
  $\xi_l^r \leftarrow \sum_{l=1}^{m} (N_l/N) \times \xi_l^r$
end for
Update global model parameter $\theta \leftarrow \xi$
Send θ to unlabeled local clients $\{C_u\}$ for initialization
for each client C_u **in parallel do** // FDST at Unlabeled Clients
  $\hat{y}_i^u \leftarrow f_\theta(x_i^u), x_i^u \in D^u$ ▷(2)
  $M_i^u \leftarrow$ uncertainty&prototype_estimation$(\hat{y}_i^u)$ ▷(4)
  $\tilde{y}_i^u \leftarrow M_i^u \cdot \hat{y}_i^u$ ▷(5)
for each round $s = 1, 2, \ldots, S$ **do**
  for each client C_u **in parallel do**
    $\psi_u^s \leftarrow$ **ClientUpdate**($\tilde{D}^u = \{(x_I^u, \tilde{y}_i^u)\}_{i=1}^{N_u}$, ψ_u^{s-1})
  end for
  $\psi_u^s \leftarrow \sum_{u=m+1}^{m+n} (N_u/N) \times \psi_u^s$
end for
for each round $t = 1, 2, \ldots, T$ **do** // FSSL at All Clients
  for each client C_l (or C_u) **in parallel do**
    ξ_l^t (or ψ_u^t) $\leftarrow$ **ClientUpdate**(D^l (or $\tilde{D}^u$), ξ_l^{t-1}(or ψ_u^{t-1}))
  end for
  $\xi_l^t, \psi_u^t \leftarrow \mu \sum_{l\in[1,m]} \eta_l \xi_l^t + (1-\mu) \sum_{u\in[m+1,m+n]} \eta_u \psi_u^t$ ▷ (7)
end for
Jump to Step 10 for pseudo-label updating (optional)

pseudo-labeling method for source-free domain adaption [20], we further extend it to our federated learning scenario to filter out or refine unreliable pseudo labels for unlabeled client models. To mitigate the unreliable areas in pseudo labels, uncertainty estimation is first performed for each image x_i^u in the client C_u using Monte Carlo Dropout [39], which enables K stochastic forward predictions through the FSL model and outputs corresponding probability results $p_{i(k)}^u = f_\xi(x_i^u)$, $k = 1,2,\ldots,K$. Then the pseudo-label denoising mask M_i^u could be determined by the uncertainty map $\sigma_i^u = std(p_{i(1)}^u, p_{i(2)}^u, \ldots, p_{i(K)}^u)$ along with an uncertainty threshold γ, namely $M_i^u = I_A(\sigma_i^u < \gamma)$. In addition, prototype estimation is further utilized to refine the result by calculating the class-wise relative feature distance map d_i^w with respect to the feature centroids t_i^w in term of class w (1: object foreground or 2: background), shown as follows:

$$t^w = \frac{\sum_v \bar{e}_v b_v^w p_v^u}{\sum_v b_v^w p_v^u} \tag{3}$$

where $\bar{e}_v$ is the average feature map of K stochastic forward outputs from the layer before the last convolution, bilinearly interpolated to be consistent with $\hat{y}_i^u$ in term of dimension. b_v^w is the binary mask computed based on the uncertainty-guided pseudo label, namely $b_v^w = I_A(\sigma_i^u < \gamma) I_A(\hat{y}_i^u = w)$, and p_v^u is the prediction probability to weigh the contributions in each pixel v. Notably, the feature centroid t_i^w is computed for pseudo-label denoising across a specific image batch each time considering suitable feature representation and computing capability. Then the relative feature distance can be easily obtained with $d_{iv}^w = \|\bar{e}_v - t^w\|$, and an updated denoising mask can be obtained, shown as follow:

$$\begin{aligned} M_i^u = {} & I_A(\sigma_i^u < \gamma) I_A(\hat{y}_i^u = 1) I_A(d_i^1 < d_i^2) \\ & + I_A(\sigma_i^u < \gamma) I_A(\hat{y}_i^u = 0) I_A(d_i^1 > d_i^2) \end{aligned} \tag{4}$$

Finally, the denoised pseudo label $\tilde{y}_{ic}^u$ could be acquired by filtering the noisy federated pseudo label with the generated mask, shown as follow:

$$\tilde{y}_i^u = M_i^u \cdot \hat{y}_i^u \tag{5}$$

3) FDST Model Generation for Unlabeled Clients: With the generated pseudo labels obtained from (4), we can successfully adapt our well-trained FSL model to each unlabeled client through self-training. To encourage the robustness of unlabeled client models, weakly random but realistic augmentation $\pi(\cdot)$ is performed on unlabeled input data, such as rotation, flipping, Gaussian noise, etc. Similar to labeled clients, the cross-entropy loss function is exploited here to minimize the discrepancy between the denoised pseudo label $\tilde{y}_i^u$ and the prediction with augmented input $\pi(x_i^u)$ for an unlabeled client C_u, shown as follow:

$$\mathcal{L}_U(\psi_u) = CrossEntropy(\tilde{y}_i^u, f_{\psi_u}(\pi(x_i^u))) \tag{6}$$

where the parameter ψ_u in each local model f_{ψ_u} is the optimization target.

C. Federated Model at Both Labeled and Unlabeled Clients

With the proposed federated pseudo-labeling strategy, the limited labeled clients can be combined with the pseudo-labeled clients to improve our FSSL model. After training with multiple epochs, the server can update the FSSL model via simply aggregating the parameters from both FSL model f_ξ and FDST model f_ψ, shown as follow:

$$\begin{aligned} \theta \leftarrow \mu\xi + (1-\mu)\psi = {} & \mu \sum\nolimits_{l\in[1,m]} \eta_l \xi_l \\ & + (1-\mu) \sum\nolimits_{u\in[m+1,m+n]} \eta_u \psi_u \end{aligned} \tag{7}$$

where ξ_l and ψ_u represent the local model parameters for labeled client C_l and unlabeled client C_u respectively, and $\eta_t(t \in [1, m+n])$ is the hyperparameter proportional to the corresponding client C_t. Particularly, considering the potential quality discrepancy of pseudo labels apart from the data quantity, treating expert annotations and pseudo labels equally is not reasonable during training. Therefore, we introduce a hyper-parameter μ in (6) to further adjust the aggregation contributions from labeled and unlabeled clients. Then, each client collects the updated FSSL model parameters from the central server and fine-tunes the model on the corresponding local dataset. Moreover, the training procedure alternates between updating the segmentation model and refining the generated pseudo labels. Specifically, the updated FSSL model is reapplied to the unlabeled clients instead of the FSL model mentioned in Section III-B to produce refined pseudo labels, which can be incorporated into the following training rounds for further improvement, namely pseudo-label updating.

IV. Experiments and Evaluation

A. Datasets and Evaluation Metrics

Experiments have been conducted to evaluate the performance of different methods with suitable metrics on both fundus image segmentation and prostate MRI segmentation tasks, where much more multi-source local datasets with cross-domain shifts are involved to well support the claimed FSSL setting, against [36] simply with only two COVID datasets utilized for the federated paradigm. In both tasks, we randomly set the split ratio of labeled to unlabeled sites to investigate the segmentation performances.

1) Fundus Image Segmentation: We validated our method on retinal fundus images collected from four different public datasets for optic disc and cup segmentation, including RIM-ONE-r3 dataset (Site A) [40], Drishti-GS dataset (Site B) [41] and another two from REFUGE challenge (Site C and D) [42]. Distribution shifts exist across those datasets due to heterogeneous imaging conditions from different clinical centers. Each retinal fundus image was cropped to a suitable target disc region with a size of 512×512 pixels as our network input.

2) Prostate MRI Segmentation: We further collected prostate T2-weighted MRI from six different data sources out of three public datasets for model validation, including NCI-ISBI 2013 dataset (Site A and B) [43], I2CVB dataset (Site C) [44] and PROMISE12 dataset (Site D, E and F) [45]. We sampled 2D slices from those 3D MRI volumes with similar field of view for the prostate region and resized them to 384×384 in axial plane.

The details of the data utilized in our experiments are summarized in Fig. 2.

3) Evaluation Metrics: In our medical image segmentation scenarios, the evaluation process is performed between the ground truth mask annotated by experts and the segmentation result predicted by a specific algorithmic model. We adopt the widely used Dice coefficient and Average Symmetric Surface Distance (ASSD) as the metrics to evaluate the segmentation performance, shown as follows:

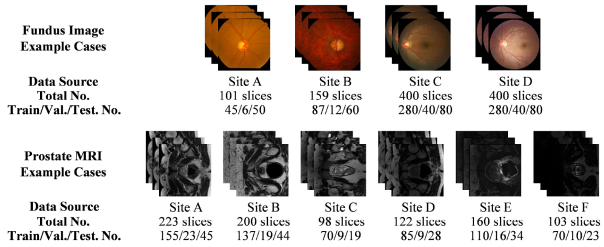

Fig. 2. Dataset details for our fundus image and prostate MRI segmentation tasks.

$$Dice = 2 \cdot \frac{|P \cap G|}{|P| + |G|} \tag{8}$$

$$ASSD(P,G) = \frac{\sum_{p \in \partial P} \min_{g \in \partial G} \|p - g\| + \sum_{g \in \partial G} \min_{p \in \partial P} \|g - p\|}{|\partial P| + |\partial G|} \tag{9}$$

where P and G mean the predicted segmentation mask and the ground truth mask, and ∂P and ∂G represent the boundary or surface of P and G.

B. Training Procedures and Implementation Details

In the FSSL procedure, a MobileNetV2 adapted DeepLabv3+ [46] is exploited as the network backbone for each local client. It is trained with Adam optimizer, and the batch size and the learning rate are set to 5 and 1×e-3 respectively. We empirically set the hyper-parameter μ to 0.58 to balance the contributions between labeled and unlabeled clients. All the unlabeled clients are initialized with the well-trained FSL models trained with 100 federated rounds. Then we sequentially perform FSSL training with generated federated pseudo labels till the global model converges (100 federated rounds), with the local epoch set to 1 in each federated round. For the federated pseudo label generation, the Monte Carlo Dropout rate of the uncertainty estimation is set to 0.5 with a stochastic forward pass time $K = 10$. The thresholds λ and γ are set to 0.75 and 0.05 respectively to guarantee the reliability of refined pseudo labels. Moreover, the federated pseudo-labeling strategy is performed every 50 federated rounds to refine the pseudo labels with the updated FSSL model. Our proposed scheme was implemented using PyTorch v1.6.0 on an NVIDIA TITAN RTX GPU.

C. Comparison With the State-of-The-Art

To evaluate the prediction performance, we first compare our approach with two well-performing unsupervised domain adaptation (UDA) methods, including **BEAL** [47]: an adversarial learning method to encourage the boundary prediction and mitigate domain shifts across different medical image datasets; and **SFDA** [20]: a source-free UDA method that leverages pseudo-label denoising scheme to promote model self-adaptation. For fair comparison, one site (Fundus Site C and Prostate Site D in our experiments) is randomly selected as labeled client and the

TABLE I
QUANTITATIVE COMPARISONS TO THE STATE-OF-THE-ART ON RETINAL FUNDUS IMAGE SEGMENTATION

Method	Client No.		Optic Cup Segmentation		Optic Disc Segmentation	
	Labeled	Unlabeled	Dice [%] ↑	ASSD [pixel] ↓	Dice [%] ↑	ASSD [pixel] ↓
FedAvg (LB)	1	3	80.04±18.58	9.00±7.85	89.54±10.94	14.81±18.77
	2	2	80.13±14.46	9.46±8.14	91.96±7.19	7.38±7.00
FedAvg (UB)	4	0	85.58±9.67	7.05±6.77	93.89±3.56	5.51±3.36
FL-CWT	4	0	82.87±10.8	7.49±5.16	93.27±4.35	5.89±3.81
BEAL	1	3	82.61±13.36	8.10±5.77	93.05±3.69	8.83±3.04
SFDA	1	3	81.15±12.30	9.44±6.85	**94.06±4.33**	7.40±8.66
Fed-ST	1	3	81.32±14.36	9.30±10.66	89.26±7.05	11.59±10.78
	2	2	82.05±15.75	8.29±8.59	92.17±5.69	7.45±6.99
FSSL-DPL (ours)	1	3	83.12±14.01	7.75±7.30	91.37±5.40	8.38±6.47
	2	2	**85.03±11.16**	**7.19±7.04**	93.42±3.59	**6.31±5.22**

The bold entities just indicate or highlight the highest performance during comparison among different methods.

rest are treated as unlabeled clients. Moreover, we compare our approach with state-of-the-art FSSL models with the same local network backbone, such as **Fed-ST** (Federated Self-Training method) [36], which employs the predicted plain pseudo-labels at unlabeled clients to update the model parameters with a consistency regularization strategy; and **FedAvg** [38], an FL framework learning from only labeled clients or from all clients fully assigned with ground truth labels, serving as the Lower-performance Bound (LB) and Upper-performance Bound (UB) in FSSL. In addition, we further exploit a typical FL method, namely CWT (Cyclic Weight Transfer) [48] in a supervised setting for comparison, which trains local clients in a serial and cyclical way. Apart from the split ratio of data sites mentioned in UDA setting, we further randomly treated two sites (Fundus Site A and C and Prostate Site D and F in our experiments) as labeled clients and the others as unlabeled clients for detailed analysis in both tasks, given the general situation where usually $|D_L| \leq |D_U|$. Additional analysis with different split ratios will be presented in Section IV-D, with Fig. 7 for illustration.

The quantitative comparison results for retinal fundus image segmentation are shown in Table I. As observed, the two UDA methods can achieve obviously better performance over FedAvg(LB) supervised with partially labeled clients, and produce comparable or slightly better results against Fed-ST utilizing plain pseudo labels (Optic Cup: BEAL 82.61% SFDA 81.15% vs. Fed-ST 81.32%/82.05%; Optic Disc: BEAL 93.05% SFDA 94.06% vs. Fed-ST 89.26%/92.17%), mainly attributing to their regularization effect on relieving across-domain biases. However, their adaptation strategies based on the individual data distributions fail to make full use of abundant multi-source data with diverse distributions to learn domain-invariance features. In contrast, our FSSL-DPL method could take better advantage of unlabeled information and generally improve the performance over the two UDA methods with much higher Dice score and lower ASSD. Specifically, without directly accessing multi-site data for domain adaptation in view of privacy protection, our method still achieves 89.23% = (85.03% + 93.42%)/2 average Dice score for optic cup and disc segmentation, which is significantly higher than BEAL with 87.83% = (82.61% + 93.05%)/2 average Dice Score and SFDA with 87.61% = (81.15% + 94.06%)/2 average Dice Score. In addition, our method has shown distinct improvements over FL-CWT even in a totally supervised setting, with Dice score increased by 2.16% = 85.03% − 82.87% and 0.15% = 93.42% − 93.27% for optic cup segmentation and disk segmentation respectively. Furthermore, our method has manifestly surpassed Fed-ST especially on optic cup segmentation, with Dice score increased by 2.98% = 85.03% − 82.05%, and is quite close to the UB results of FedAvg with a supervised setting when $|D_L|:|D_U|$ = 2:2. These results confirm the superiority of our method by jointly exploiting both labeled and unlabeled clients with our federated pseudo-labeling strategy to effectively filter out unreliable pseudo labels and boost the distributed learning over medical data at multiple sites. Fig. 3 illustrates the qualitative comparison of different methods in the retinal fundus image segmentation task, which intuitively shows that our method can provide more accurate segmentation results with clear prediction entropy maps over multi-site datasets with domain shifts.

For prostate MRI segmentation, we could find similar results compared with fundus image segmentation, as shown in Table III. Specifically, SFDA with 88.99% Dice score obtains more than 2% Dice improvement over FedAvg(LB) (86.14%/87.30% Dice score) and achieves comparable or slightly worse results than Fed-ST (89.21%/90.15% Dice score). Our FSSL-DPL significantly outperforms all the UDA methods and exceeds Fed-ST with improvement from 90.15% to 91.95% in Dice and from 5.32 to 4.27 in ASSD when $|D_L|:|D_U| = 2:4$. Remarkably, our method even outperforms the supervised FL method FL-CWT (91.95% vs. 89.26%) and shows the closest performance to the upper bound given by FedAvg. Somewhat differently, BEAL shows a relatively poor performance in this case, even worse than FedAvg(LB), which may imply the possible limitation of BEAL when dealing with different multi-site scenarios. Qualitative comparison examples for the prostate MRI segmentation task are shown in Fig. 4. We can observe that our method can implement accurate segmentation with more certain (low-entropy) predictions across all the six data sites, whereas the other methods sometimes fail when facing distribution gap.

To further demonstrate the advantages of our method, we perform t-test in terms of Dice score to show the significance level against different methods, as shown in Table V. We can find

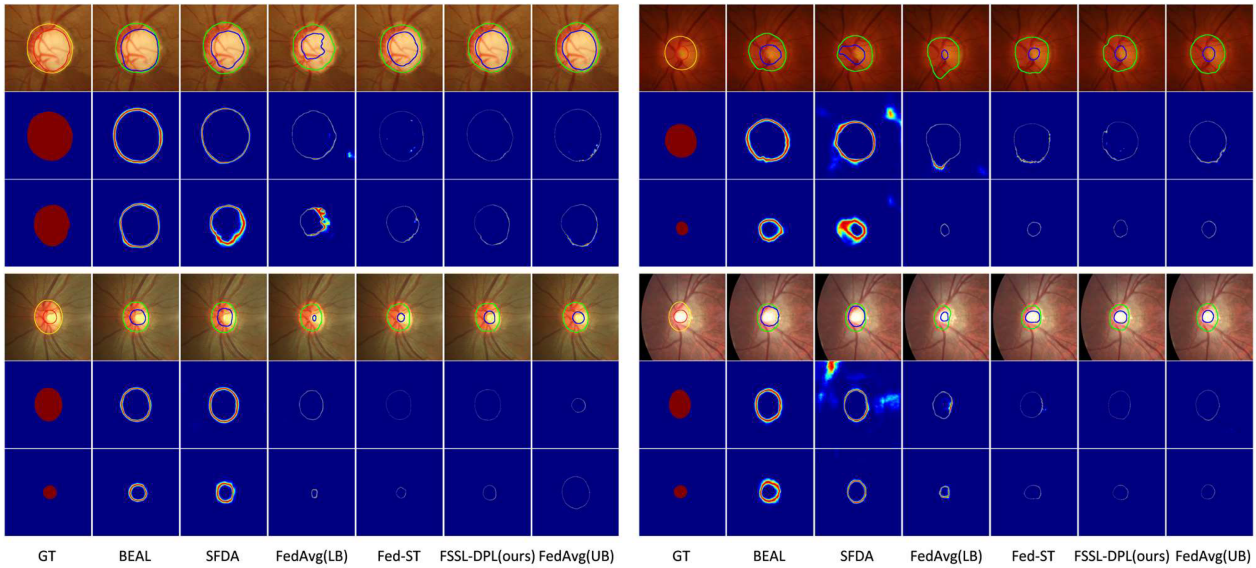

Fig. 3. Qualitative comparison examples in the retinal fundus image segmentation task. For each domain-specific example, the segmentation results for optic disk (green) and cup (blue) are marked with boundary lines. The segmentation entropy maps rescaled to [0, 1] for better visualization are also presented to indicate the prediction uncertainty, where red color means high entropy values.

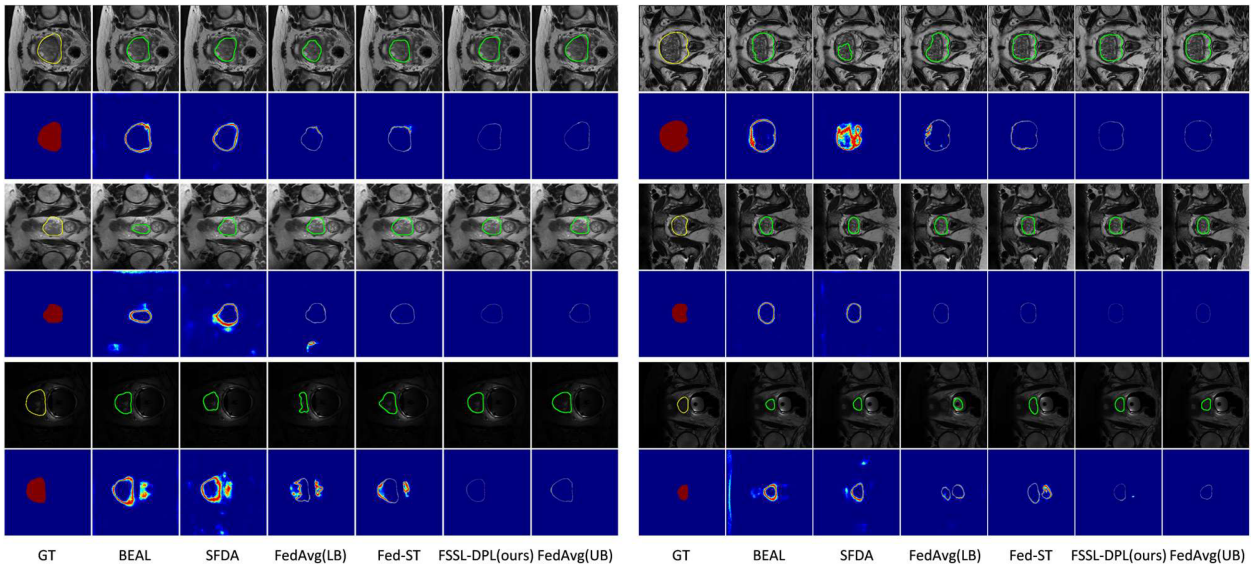

Fig. 4. Qualitative comparison examples in the prostate MRI segmentation task. For each domain-specific example, the prostate segmentation results are marked with green boundary lines. The segmentation entropy maps rescaled to [0, 1] for better visualization are also presented to indicate the prediction uncertainty, where red color means high entropy values.

that for fundus cup segmentation, t(BEAL), t(SFDA) and t(Fed-ST) $>$ $t_{0.05/269}$ while t(FedAvg(UB)) $<$ $t_{0.05/269}$ [$t_{0.05/200}$ = 1.653, $t_{0.05/500}$ = 1.648], thus we know there are significant difference between BEAL, SFDA, Fed-ST and our method, and our performance is quite close to the upper bound FedAvg(UB). For fundus disk segmentation, only t(Fed-ST) $>$ $t_{0.05/269}$, which means that our method is better than Fed-ST and has no significant difference against the other methods. Similarly, we can also find that t(BEAL), t(SFDA) and t(Fed-ST) $>$ $t_{0.01/192}$ while t(FedAvg(UB)) $<$ $t_{0.01/192}$ [$t_{0.05/100}$ = 2.626, $t_{0.01/200}$ = 2.601], thus we know there are significant difference between BEAL, SFDA, Fed-ST and our method, and our performance is close to the upper bound FedAvg (UB). Generally, based on above t-test analysis, we have validated that our method is much more effective against the other methods.

D. Ablation Study

1) Contribution of Each Component: We performed an ablation analysis to investigate the effectiveness of our federated pseudo-labeling (PL) strategy including PL denoising and PL

TABLE II
ABLATION STUDY IN THE RETINAL FUNDUS IMAGE SEGMENTATION TASK

Method	Dice [%] ↑			ASSD [pixel] ↓		
	Optic Cup	Optic Disc	Avg.	Optic Cup	Optic Disc	Avg.
Fed-ST (baseline)	82.05	92.17	87.11	8.29	7.45	7.87
+ PL Denosing	83.89	92.14	88.02	8.06	7.86	7.96
+ PL Updating	84.71	93.13	88.92	**6.69**	6.72	**6.71**
+ Hyperparam. Tuning	83.87	93.20	88.54	7.78	6.55	7.17
FSSL-DPL(ours)	**85.03**	**93.42**	**89.23**	7.19	**6.31**	6.75

The bold entities just indicate or highlight the highest performance during comparison among different methods.

TABLE III
QUANTITATIVE COMPARISONS TO THE STATE-OF-THE-ART ON PROSTATE MRI SEGMENTATION

Method	Client No.		Prostate Segmentation	
	Labeled	Unlabeled	Dice [%] ↑	ASSD [pixel] ↓
FedAvg (LB)	1	3	86.14±12.64	10.01±12.67
	2	4	87.30±13.99	7.16±8.53
FedAvg (UB)	6	0	93.15±4.87	3.38±2.18
FL-CWT	6	0	89.26±11.24	4.96±4.60
BEAL	1	5	80.50±12.63	10.32±5.03
SFDA	1	5	88.99±4.95	6.31±6.40
Fed-ST	1	3	89.21±12.39	5.67±6.09
	2	4	90.15±6.50	5.32±3.51
FSSL-DPL (ours)	1	3	90.20±9.33	5.61±5.26
	2	4	**91.95±4.54**	**4.27±3.83**

The bold entities just indicate or highlight the highest performance during comparison among different methods.

TABLE IV
ABLATION STUDY IN THE PROSTATE MRI SEGMENTATION TASK

Method	Dice [%] ↑	ASSD [pixel] ↓
Fed-ST (baseline)	90.15	5.32
+ PL Denoising	91.52	4.40
+ PL Updating	91.77	**4.21**
+ Hyperparam. Tuning	91.12	4.43
FSSL-DPL(ours)	**91.95**	4.27

The bold entities just indicate or highlight the highest performance during comparison among different methods.

TABLE V
T-TEST FOR SIGNIFICANCE LEVEL ANALYSIS

Fundus Seg.	BEAL	SFDA	Fed-ST	FedAvg (UB)
t val. (cup/disk)	2.28/1.18	3.83/1.28	2.53/3.04	0.61/1.52
Prostate Seg.	BEAL	SFDA	Fed-ST	FedAvg (UB)
t val.	11.82	6.11	3.16	2.49

The bold entities just indicate or highlight the highest performance during comparison among different methods.

updating and the hyperparameter tuning in aggregation contribution adjustment from labeled and unlabeled clients. We treat the Fed-ST setting as the baseline, then we analyze the effectiveness of our PL strategy by + PL Denoising and + PL Updating incrementally. Besides, we independently analyze the effect of the hyperparameter μ by + Hyperparam. Tuning based on Fed-ST. Combining all the components mentioned above, our proposed approach FSSL-DPL is obtained. The quantitative comparison results of different algorithm components are shown in Tables II and IV where $|D_L| = 2$. We can find each component can independently improve the prediction accuracy compared to Fed-ST (baseline) simply with a plain pseudo-labeling strategy, and their integration can further improve the fundus image and prostate MRI segmentation performances in terms of Dice score on the entire datasets. The results show complementary role of each component contributing effectively to the performance improvement of our approach. Specifically, the pseudo-label denoising strategy lays better learning foundation with refined pseudo labels for the contribution adjustment by tuning hyperparameter between labeled and unlabeled clients. Inversely, the contribution adjustment also makes better use of the unlabeled information inside the whole FSSL system. The experimental evidence coincides with the consensus that more accurate segmentation annotations lead to a better prediction model, and also reflects the rationale of our approach to give higher aggregation weight to labeled client sides at the training stage. In addition, Fig. 5 shows detailed examples of our pseudo-label denoising strategy in the training procedure. As the training goes on, our FSSL model iteratively updates the model parameters and the predicted pseudo labels of the unlabeled clients, which could yield further model improvements along with refined pseudo labels that are much closer to ground truth compared to the plain pseudo labels used in Fed-ST. Furthermore, we analyzed the specific performance of our method on each individual client/site compared with the other two federated methods FedAvg(LB) and Fed-ST, as shown in Fig. 6, which reflects that our method yields comparable or better performance than the other two methods, especially on most unlabeled sites with clearly higher margins, e.g., FSSL-DPL(ours) 90.67% > Fed-ST 86.11% > FegAvg(LB) 85.87% on Fundus Site D, and FSSL-DPL(ours) 93.7% > Fed-ST 88.65% > FegAvg(LB) 85.57% on Prostate Site E. Notably, there is a performance decrease on unlabeled Site B compared with FedAvg(LB) for fundus image segmentation. This is mainly because the FedAvg(LB) model learned on labeled Site A and C fails to learn a relatively generalized model due to inter-client interference, which tends to converge to Site A and consequently leads to a fairly low results on Site C. The distribution of unlabeled Site B coincidently close to that of Site A can obtain high performance. In contrast, our model can effectively learn generalizability in complex distribution

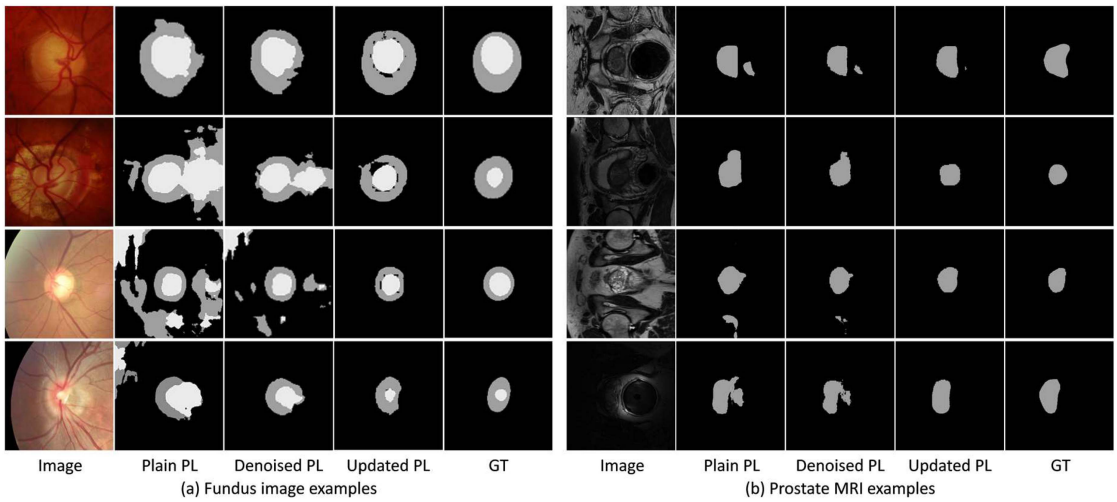

Fig. 5. Examples of our federated pseudo-labeling strategy by iteratively updating the model parameters and the predicted pseudo labels (PLs) of unlabeled clients.

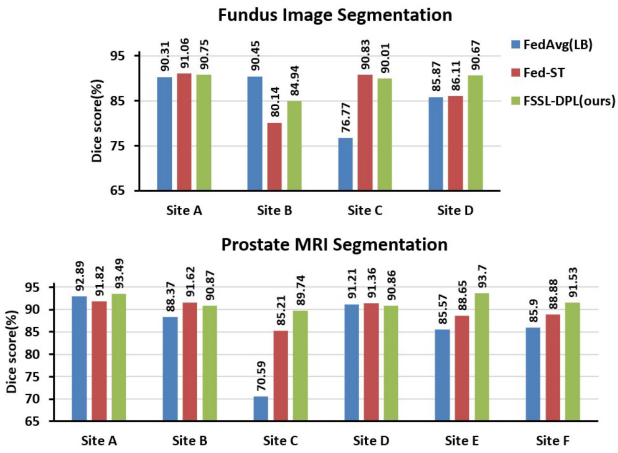

Fig. 6. Comparison results of different federated algorithms in each individual site to analyze the effect of our federated pseudo-labeling strategy.

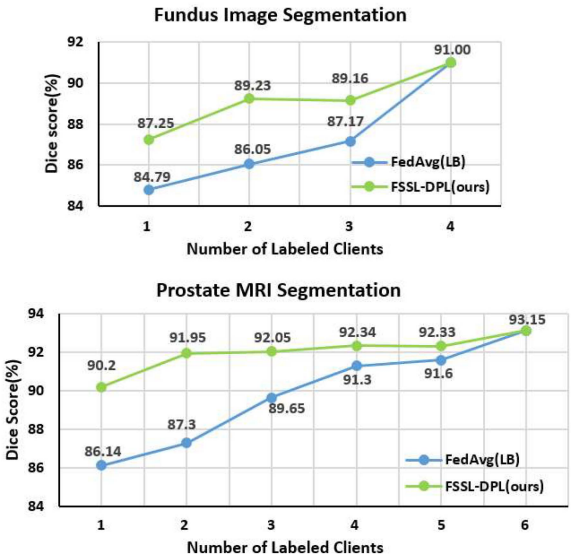

Fig. 7. Curves of segmentation performance on fundus image and prostate MRI datasets respectively as the number of participating labeled clients increases up to the same supervised settings, using FSSL-DPL(ours) and FedAvg.

heterogeneity and produce relatively high and smooth prediction results across different data sources.

2) Effect of Participating Labeled Client Number: To further understand the relationship between model performance and participating numbers of labeled clients in FL, we conducted a series of comparative experiments between FSSL-DPL and FedAvg as the labeled client number gradually increases from 1 to $m+n$. Fig. 7 illustrates the segmentation performance on fundus and prostate MRI segmentation tasks respectively considering different number of participating labeled clients. For both methods, the segmentation performances show a constantly increasing trend as more labeled clients are involved in the federated training, demonstrating the reasonable expectation that more comprehensive data distribution can boost the learning effect in federated paradigm. Meanwhile, we observe that our FSSL-DPL method consistently surpasses FedAvg on all the semi-supervised settings with different labeled client numbers, demonstrating the effectiveness of our federated pseudo-label denoising strategy to leverage all the unlabeled information. Particularly, when the number of labeled clients reaches 2, our method can already achieve relatively high and acceptable segmentation results on both tasks against FegAvg, namely $89.23\% > 86.05\%$ on fundus segmentation task and $91.95\% > 87.3\%$ on prostate MRI segmentation task, which amply indicates its efficiency and robustness. This advantage may greatly reduce the labeling burden of medical institutions in practice and facilitate multi-institutional collaborations with FL process.

V. Conclusion

In this article, we propose a novel FSSL method for multi-site medical image segmentation, which can effectively exploit the

valuable information embedded in extensive unlabeled clients to boost the generalizability and robustness of our federated model while preserving cross-site data privacy as well. To mitigate the inevitable domain shifts between labeled and unlabeled domains, a federated pseudo-labeling strategy is introduced to generate reliable pseudo labels to train the local models from unlabeled clients, assisted by the well-trained FSL model with wider knowledge generalization learned from labeled domains. Experiments on fundus image and prostate MRI segmentation tasks with comprehensive ablation studies have successfully demonstrated the effectiveness of our approach compared with the state-of-the-art, achieving the highest Dice scores of 89.23% and 91.95% respectively even with only a few labeled clients participating in model training. Still, it is meaningful to further explore the generalizability of our method on other medical segmentation tasks considering heterogeneous anatomical regions, such as the small pancreas segmentation with complex structure and ambiguous boundary in the CT-scanned abdominal images, and the brain tumor segmentation with irregular shapes and uncertain locations. Our method has the potential to improve the accuracy and efficiency of diagnostic models, develop more personalized treatment plans and advance medical research by exploiting large multi-institution data even with limited segmentation labels. In addition, there are other challenging open questions in this research field, such as more severe cross-domain shifts among different federated clients, the potential effects caused by unbalanced ratio of labeled to unlabeled data when a larger number of clients are involved and the biased model aggregation issue due to quality discrepancies of local models. In the future, we will incorporate more sophisticated pseudo-labeling strategies to improve our scheme, such as active learning, transfer learning, uncertainty-aware rectification and domain adaptation techniques. Furthermore, more advanced FL algorithms will also be investigated to enhance the performance of our FSSL system.

REFERENCES

[1] S. C. Mathews, M. J. McShea, C. L. Hanley, A. Ravitz, A. B. Labrique, and A. B. Cohen, "Digital health: A path to validation," *NPJ Digit. Med.,* vol. 2, no. 1, 2019, Art. no. 38.

[2] C. Guo, H. Ashrafian, S. Ghafur, G. Fontana, C. Gardner, and M. Prime, "Challenges for the evaluation of digital health solutions—A call for innovative evidence generation approaches," *NPJ Digit. Med.,* vol. 3, no. 1, 2020, Art. no. 110.

[3] D. V. Gunasekeran, R. M. W. W. Tseng, Y.-C. Tham, and T. Y. Wong, "Applications of digital health for public health responses to COVID-19: A systematic scoping review of artificial intelligence, telehealth and related technologies," *NPJ Digit. Med.,* vol. 4, no. 1, 2021, Art. no. 40.

[4] G. Fagherazzi, C. Goetzinger, M. A. Rashid, G. A. Aguayo, and L. Huiart, "Digital health strategies to fight COVID-19 worldwide: Challenges, recommendations, and a call for papers," *J. Med. Internet Res.,* vol. 22, no. 6, 2020, Art. no. e19284.

[5] L. Qiu and H. Ren, "RSegNet: A joint learning framework for deformable registration and segmentation," *IEEE Trans. Automat. Sci. Eng.*, vol. 19, no. 3, pp. 2499–2513, Jul. 2022.

[6] L. Qiu and H. Ren, "U-RSNet: An unsupervised probabilistic model for joint registration and segmentation," *Neurocomputing*, vol. 450, pp. 264–274, 2021.

[7] L. Qiu, C. Li, and H. Ren, "Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural network," *Healthcare Technol. Lett.*, vol. 6, no. 6, pp. 159–164, 2019.

[8] H. Gao et al., "SAVAnet: Surgical action-driven visual attention network for autonomous endoscope control," *IEEE Trans. Automat. Sci. Eng.*, early access, Sep. 19, 2022, doi: 10.1109/TASE.2022.3203631.

[9] H. Gao, X. Xiao, L. Qiu, M. Q.-H. Meng, N. K. K. King, and H. Ren, "Remote-center-of-motion recommendation toward brain needle intervention using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 8295–8301.

[10] N. Peiffer-Smadja, R. Maatoug, F.-X. Lescure, E. D'ortenzio, J. Pineau, and J.-R. King, "Machine learning for COVID-19 needs global collaboration and data-sharing," *Nature Mach. Intell.,* vol. 2, no. 6, pp. 293–294, 2020.

[11] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.,* vol. 37, no. 3, pp. 50–60, May 2020.

[12] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–16.

[13] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 31, no. 9, pp. 3400–3413, Sep. 2020.

[14] Q. Dou et al., "Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study," *NPJ Digit. Med.,* vol. 4, no. 1, 2021, Art. no. 60.

[15] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.,* vol. 2, no. 6, pp. 305–311, 2020.

[16] N. Rieke et al., "The future of digital health with federated learning," *NPJ Digit. Med.,* vol. 3, no. 1, 2020, Art. no. 119.

[17] O. Aouedi, A. Sacco, K. Piamrat, and G. Marchetto, "Handling privacy-sensitive medical data with federated learning: Challenges and future directions," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 2, pp. 790–803, Feb. 2023, doi: 10.1109/JBHI.2022.3185673.

[18] L. Qiu and H. Ren, "U-RSNet: An unsupervised probabilistic model for joint registration and segmentation," *Neurocomputing,* vol. 450, pp. 264–274, 2021.

[19] L. Zhu, K. Yang, M. Zhang, L. L. Chan, T. K. Ng, and B. C. Ooi, "Semi-supervised unpaired multi-modal learning for label-efficient medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 394–404.

[20] C. Chen, Q. Liu, Y. Jin, Q. Dou, and P.-A. Heng, "Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2021, pp. 225–235.

[21] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*, 2018, pp. 92–104.

[22] W. Li et al., "Privacy-preserving federated brain tumour segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2019, pp. 133–141.

[23] X. Gong et al., "Ensemble attention distillation for privacy-preserving federated learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15076–15086.

[24] I. Dayan et al., "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Med.,* vol. 27, no. 10, pp. 1735–1743, 2021.

[25] M. J. Sheller et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[26] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med. Image Anal.,* vol. 65, 2020, Art. no. 101765.

[27] P. Guo, P. Wang, J. Zhou, S. Jiang, and V. M. Patel, "Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2423–2432.

[28] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1013–1023.

[29] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12674–12684.

[30] X. Li et al., "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[31] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. De Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2019, pp. 810–818.

[32] C. Li et al., "Self-ensembling co-training framework for semi-supervised COVID-19 CT segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 11, pp. 4140–4151, Nov. 2021.

[33] W. Jeong, J. Yoon, E. Yang, and S. J. Hwang, "Federated semi-supervised learning with inter-client consistency & disjoint learning," in *Proc. Int. Conf. Learn. Representations*, 2021.

[34] N. Dong and I. Voiculescu, "Federated contrastive learning for decentralized unlabeled medical images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 378–387.

[35] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu, "Federated contrastive learning for volumetric medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2021, pp. 367–377.

[36] D. Yang et al., "Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan," *Med. Image Anal.*, vol. 70, 2021, Art. no. 101992.

[37] Q. Liu, H. Yang, Q. Dou, and P.-A. Heng, "Federated Semi-supervised medical image classification via inter-client relation matching," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2021, pp. 325–335.

[38] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[39] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[40] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "RIM-ONE: An open retinal image database for optic nerve evaluation," in *Proc. IEEE 24th Int. Symp. Comput.-Based Med. Syst.*, 2011, pp. 1–6.

[41] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, and A. S. Tabish, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomed. Imag. Data Papers,* vol. 2, no. 1, 2015, Art. no. 1004.

[42] J. I. Orlando et al., "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Med. Image Anal.*, vol. 59, 2020, Art. no. 101570.

[43] N. Bloch et al., "NCI-ISBI 2013 challenge: Automated segmentation of prostate structures," *Cancer Imag. Arch.*, vol. 370, no. 6, pp. 5, 2015.

[44] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review," *Comput. Biol. Med.,* vol. 60, pp. 8–31, 2015.

[45] G. Litjens et al., "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Med. Image Anal.,* vol. 18, no. 2, pp. 359–373, 2014.

[46] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis*, 2018, pp. 801–818.

[47] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "Boundary and entropy-driven adversarial learning for fundus image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 102–110.

[48] K. Chang et al., "Distributed deep learning networks among institutions for medical imaging," *J. Amer. Med. Inform. Assoc.,* vol. 25, no. 8, pp. 945–954, 2018.