FedMix: Mixed Supervised Federated Learning for Medical Image Segmentation

Jeffry Wicaksana, Zengqiang Yan, *Member, IEEE*, Dong Zhang, Xijie Huang, Huimin Wu, Xin Yang, *Member, IEEE*, and Kwang-Ting Cheng, *Fellow, IEEE*

***Abstract*— The purpose of federated learning is to enable multiple clients to jointly train a machine learning model without sharing data. However, the existing methods for training an image segmentation model have been based on an unrealistic assumption that the training set for each local client is annotated in a similar fashion and thus follows the same image supervision level. To relax this assumption, in this work, we propose a label-agnostic unified federated learning framework, named FedMix, for medical image segmentation based on mixed image labels. In FedMix, each client updates the federated model by integrating and effectively making use of all available labeled data ranging from strong pixel-level labels, weak bounding box labels, to weakest image-level class labels. Based on these local models, we further propose an adaptive weight assignment procedure across local clients, where each client learns an aggregation weight during the global model update. Compared to the existing methods, FedMix not only breaks through the constraint of a single level of image supervision but also can dynamically adjust the aggregation weight of each local client, achieving rich yet discriminative feature representations. Experimental results on multiple publicly-available datasets validate that the proposed FedMix outperforms the state-of-the-art methods by a large margin. In addition, we demonstrate through experiments that FedMix is extendable to multi-class medical image segmentation and much more feasible in clinical scenarios. The code is available at: https://github.com/Jwicaksana/FedMix.**

***Index Terms*— Federated learning, mixed supervision, medical image segmentation, pseudo labeling, adaptive weight aggregation.**

Manuscript received 1 October 2022; revised 23 December 2022; accepted 28 December 2022. Date of publication 30 December 2022; date of current version 29 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62202179, Grant 62061160490, and Grant 61872417; in part by the National Natural Science Foundation of China/Research Grants Council Joint Research Scheme under Grant N_HKUST627/20; and in part by the Natural Science Foundation of Hubei Province of China under Grant 2022CFB585. *(Corresponding author: Zengqiang Yan.)*

Jeffry Wicaksana and Kwang-Ting Cheng are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: jwicaksana@connect.ust.hk; timcheng@ust.hk).

Zengqiang Yan and Xin Yang are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: z_yan@hust.edu.cn; xinyang2014@hust.edu.cn).

Dong Zhang, Xijie Huang, and Huimin Wu are with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: dongz@ust.hk; xhuangbs@connect.ust.hk; hwubl@connect.ust.hk).

Digital Object Identifier 10.1109/TMI.2022.3233405

I. Introduction

MEDICAL image segmentation is a representative task for image content analysis supporting computer aided diagnosis, which can not only recognize the lesion category but also locate the specific areas [1]. In the past few years, deep learning has dominated this task and has been applied in a wide range of underlying scenarios, *e.g.*, lung nodule segmentation [2], COVID-19 lesion segmentation [3], and skin lesion boundary detection [4].

The optimization of deep learning models usually relies on a vast amount of training data [5]. For example, for a fully-supervised semantic segmentation model, the ideal scenario is that we can collect the pixel-level annotated images as much as possible from diverse sources. However, this scenario is almost infeasible due to the following two reasons: 1) the strict sharing protocol of sensitive patient information between medical institutions and 2) the exceedingly high pixel-level annotation cost. As the expert knowledge usually required for annotating medical images is much more demanding and difficult to obtain, various medical institutions have very limited strong pixel-level annotated images and most available images are unlabeled or weakly-annotated [3], [7], [21], [22], [27]. Therefore, a realistic clinical mechanism that utilizes every available supervision for cross-institutional collaboration without data sharing is highly desirable.

Thanks to the timely emergence of Federated Learning (FL), which aims to enable multiple clients to jointly train a machine learning model without sharing data, the problem of data privacy being breached can be alleviated [12]. FL has gained significant attention in the medical imaging community [13], [18], due to the obvious reason that medical images often contain some personal information. During the training process of a standard FL model, each local client first downloads the federated model from a server and updates the model locally. Then, the locally-trained model parameters of each client are sent back to the server. Finally, all clients' model parameters are aggregated to update the global federated model. Most of the existing FL frameworks [14], [19] require that the data used for training by each local client needs to follow the same level of labels, *e.g.*, pixel-level labels (as shown in Fig. 1 (d)) for an image semantic segmentation model, which limits the model learning ability. Though some semi-supervised federated learning methods [33], [36] attempt to utilize the unlabeled data in addition to pixel-level labeled

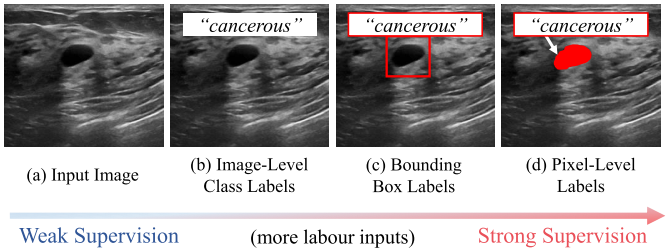

Fig. 1. Examples of different levels of medical image labels, where the image-level class label in (b) contains only the lesion category. The bounding box label in (c) contains not only the lesion category but also a coarse location. The pixel-level label in (d) contains both the lesion category and location information of each pixel, which is strong image supervision. Though strong image supervisions are more informative, they are very expensive to obtain. The utilization of some easy-to-access image supervision is beneficial in practice.

images in training, the weakly-labeled images (*e.g.*, image-level class labels in Fig. 1 (b) and bounding box labels in Fig. 1 (c)), were ignored.

Clients participating in FL may have different labeling budgets. Therefore, there may be a wide range of inter-client variations in label availability. Weak labels are easier to acquire and thus more broadly available compared to pixel-level ones. In practice, there can be various types of weak labels. While an image-level label indicating whether a breast ultrasound image is cancerous or not is easier to acquire compared to a bounding box label pointing out the specific location of the cancerous region, it is also less informative. Effectively utilizing the information from these weakly-labeled data with varying levels of label strengths as well as unlabeled data, especially for clients without pixel-level labeled data, would be highly beneficial for improving the federated model's robustness while preventing training instability. However, given less informative training data, ensuring that the local model updates from clients without pixel-level labels, *e.g.*, weakly-labeled and unlabeled clients, positively contribute to the federated model remains a challenge.

In this work, as illustrated in Fig. 2, we propose the first label-agnostic Mixed Supervised Federated Learning (FedMix) framework, which makes full use of data labeled in any form to train a unified FL model for medical image segmentation. We ensure positive contributions from every client's local model updates through a two-step process. First, in the absence of pixel-level labels, FedMix effectively utilizes consistency training to extract useful information from the unlabeled images as well as the weakly-labeled images (*i.e.*, image-level class labels and bounding box labels) for producing and selecting high-quality pseudo labels used for local model updates. Through an iterative process, the accuracy of selected pseudo labels gradually improves, leading to better local model performance. Then, to better estimate the potential contribution of each client, its training loss is used as a metric to adaptively adjust each client's aggregation weight in the update of the federated model, which is vital to handle the inter-client variations in label availability. Compared to the existing methods, FedMix not only alleviates the constraint of a single type of label but also achieves better convergence through dynamically assigning an optimized aggregation weight to each local client. Experimental results on several publicly-available datasets demonstrate the superior performance of FedMix under both semi-supervised and mixed-supervised settings, both of which are more realistic in clinical scenarios. The contributions are summarized as follows:

- The first mixed supervised FL framework for medical image segmentation through an iterative pseudo label generator followed by a label refinement operation, based on the information derived from weakly-labeled data, to target high-quality pseudo labels for training.
- Adaptive weight aggregation across clients to handle inter-client variations in supervision availability, which is proven to be extendable to other FL frameworks.
- Superior performance on the challenging breast tumor segmentation and skin lesion segmentation. FedMix outperforms the state-of-the-art methods by a large margin.

The rest of this paper is organized as follows. Existing and related work are summarized and discussed in Section II. Details of FedMix are introduced in Section III. In Section IV, we present a thorough evaluation of FedMix compared with the existing methods and provide ablation studies as well as analysis in Section V. Section VI concludes the paper.

II. Related Work

A. Federated Learning

Federated learning (FL) is a distributed learning framework, which is designed to allow different clients, institutions, and edge devices to jointly train a machine learning model without sharing the raw data [12], which plays a big role in protecting data privacy. In recent years, FL has drawn great attention from the medical image communities [19], [58] and has been validated for multi-site functional magnetic resonance imaging classification [14], health tracking through wearables [64], COVID-19 screening and lesion detection [59], and brain tumor segmentation [13], [18]. In clinical practice, different clients may have great variations in data quality, quantity, and supervision availability. Improper use of these data may lead to significant performance degradation among different clients. To reduce inter-client variations, FL has been combined with domain adaptation [17], [65], [68], contrastive learning [66] and knowledge distillation [67] to learn a more generalizable federated model. However, existing works do not consider the variations in supervision availability (*i.e.*, different clients have different levels of labels), which is commonly observed in clinical practice.

B. Federated Semi-Supervised Learning

In a standard federated learning setting, not every local client has access to pixel-level supervision for image segmentation to facilitate model learning with weakly-labeled and unlabeled training data. To this end, some federated semi-supervised learning approaches require clients to share supplementary information, *e.g.*, client-specific disease relationship [34], extracted features from raw data [37], metadata of the training data [38], and ensemble predictions from

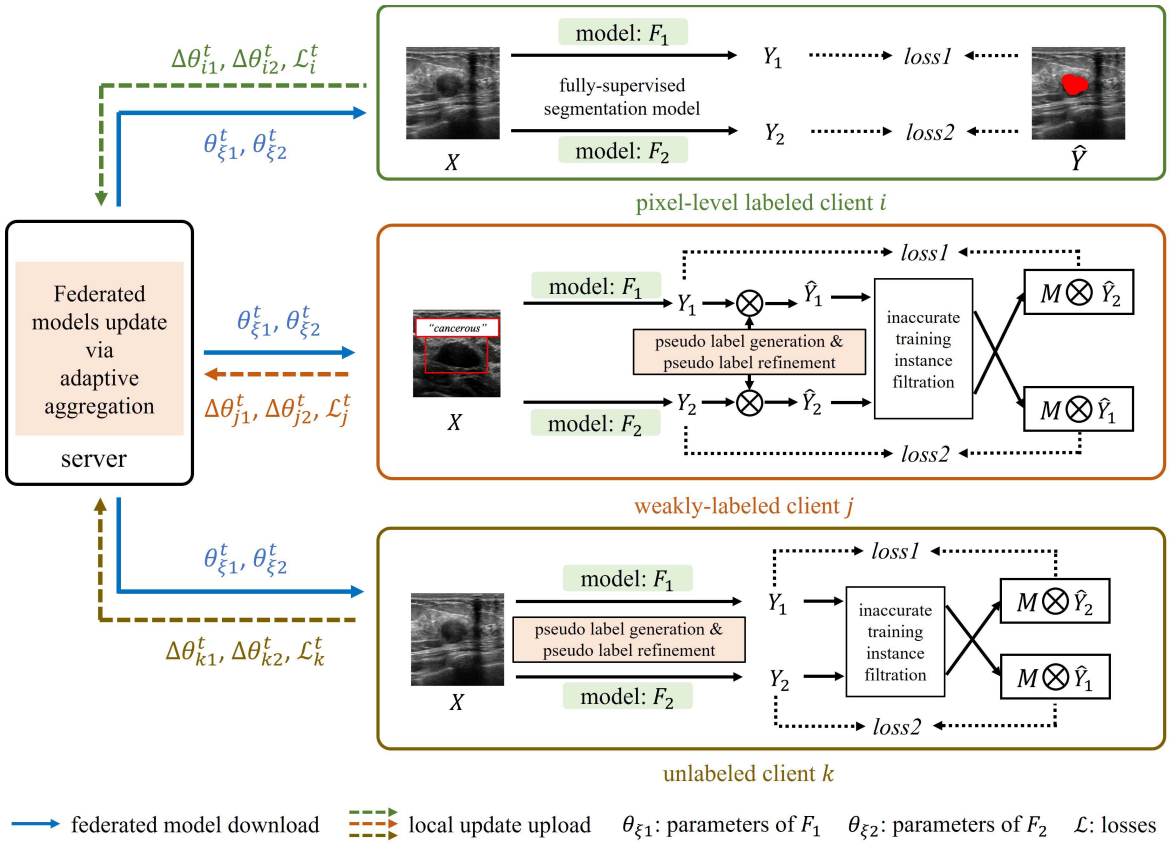

Fig. 2. Illustration of the proposed Mixed Supervised Federated Learning (FedMix) framework. The local client update utilizes every available supervision for training. Based on this, an adaptive weight aggregation procedure is used for the global federated model update. Compared to existing methods, FedMix not only breaks through the constraint of a single level of image supervision but also can dynamically adjust the aggregation weight of each local client, achieving rich yet discriminative feature representations.

different clients' locally-updated models besides their parameters [36]. FedMatch [46] attempted to utilize unlabeled data from various clients by first sharing the client-consistency matrix in the embedding space to assign various helper models for participating clients. Additional information sharing beyond the locally-updated model parameters may leak privacy-sensitive information [57] about clients' data. FedRGD [35] learned from unlabeled clients by minimizing the gradient diversity among clients through the usage of consistency training, group norm [54], and a novel federated aggregation method. Yang et al. [33] proposed to avoid additional information sharing by first training a fully-supervised federated learning model only on clients with available pixel-level supervision for several training rounds and then using the model to generate pseudo labels for local clients based on the unlabeled data. Those confident pseudo labels are used to supervise the local model updates on unlabeled clients for subsequent rounds.

C. Medical Image Segmentation

The deep learning-based image recognition technology has been used for various medical image segmentation tasks, *e.g.*, brain tumor segmentation [41], [42], optic disc segmentation [25], lung nodules segmentation [2], lesion boundary segmentation [4], and COVID-19 lesion segmentation [3]. However, training a fully-supervised deep model for image semantic segmentation often requires access to a mass of pixel-level supervisions, which are expensive to acquire [22]. In particular, the problem of the expensive pixel-level supervision is much more obstructive for medical image segmentation [28]. To this end, efforts have been made to explore the use of some easily obtained image supervisions [43], [44] (*e.g.*, scribbles [55], image-level classes [6], bounding boxes [8], points [9], and even unlabeled image [39]) to train a pixel-level image segmentation model. However, most of the existing works are based on only one or two types of image supervision, which greatly limits the model learning efficiency. In most cases, access to some pixel-level annotated data is required to facilitate model training, which may not always be available for each participating client.

III. Our Approach

In this section, we first introduce the notation and experimental settings of the proposed unified federated learning framework, *i.e.*, FedMix, in Section III-A. Then, we provide a framework overview in Section III-B. We present implementation details including pseudo label generation, selection, and federated model update of the proposed FedMix in Section III-C and Section III-D. Finally, we reformulate FedMix from different perspectives in Section III-E.

A. Preliminaries

1) Experimental Settings: To emulate real scenarios, we focus on deep learning from multi-source datasets, where clients' data is collected from different medical sources. To explore variations in cross-client supervision, each client is assumed to own only one label type.

2) Training Notations: In this paper, we denote $\overline{D} = [D_1, \ldots, D_N]$ as the collection of N clients' training data. Given client i, $D_i^L = [X, Y_{gt}]$, $D_i^U = [X]$, $D_i^{img} = [X, Y_{img}]$, and $D_i^{bbox} = [X, Y_{bbox}]$ represent the training data that is pixel-level labeled, unlabeled, image-level class labeled, and bounding box-level labeled, respectively. X and Y represent the sets of the training images and the available labels.

To integrate various levels of labels, we modify the bounding box labels and image-level class labels to pixel-level labels. Specifically, the bounding box point representation is converted into pixel-level labels where the foreground class falls inside the bounding box and the background class falls outside the bounding box. For image-level class labels, we constrain the pixel-level label to the corresponding image class. Consequently, Y_{gt}, Y_{img}, and Y_{bbox} has the same dimension, *e.g.*, $Y \in \mathbb{R}^{C \times H \times W}$, where C indicates the total number of foreground classes and $W \times H$ indicates the size of the respective image data. For $Y_{img} = \{y_1, \ldots, y_C | y_i \in [0, 1]^{H \times W}\}$, $y_i = 1$ if the class $i \in C$ is present in the image-level label.

B. Overview

Weakly-labeled and unlabeled clients may not have a sufficient amount of reliable information for model training, which can negatively affect the local model updates and thus, in turn, the federated model. To fully utilize *every level of labels* available at any client, FedMix addresses the challenge through:

1) **Pseudo Label Generation and Selection *(Sample & Refine)*.** To utilize *every available data* and ensure reliable local model updates from clients without pixel-level labels, we design a novel unified framework using *every level of label* to amplify and filter useful signals from pseudo supervision. Specifically, FedMix utilizes consistency regularization [39] to generate pseudo labels which are then dynamically filtered and refined before being used for training.
2) **Adaptive Aggregation for Federated Model Update *(Aggregate)*.** FedMix presents a novel adaptive aggregation operation to alleviate the training instability which may arise from naively aggregating local model updates from weakly-labeled and unlabeled clients. By taking consistency regularization and dynamic sample selection into account, the weight of each client is determined according to its data quantity and quality (inferred from training loss). In this way, more reliable clients will be assigned with higher weights, leading to better convergence.

We illustrate FedMix in Fig. 2 and present the pseudo-code in Algorithm 1.

Algorithm 1 Pseudocode of FedMix

input : $\overline{D}$: the set of training data
parameter: β, λ: hyperparameters for adaptive aggregation
T: maximum federated training rounds
ϵ: threshold for dynamic sample selection
output : $\theta_{\xi 1}^T$: parameters of F_1
$\theta_{\xi 2}^T$: parameters of F_2

$\theta_{\xi 1}^0, \theta_{\xi 2}^0 \leftarrow$ initialize()
for $t = 1 : T$ **do**
  $\overline{\mathcal{L}} = \{\}, \overline{\theta}_{\xi 1} = \{\}, \overline{\theta}_{\xi 2} = \{\}$
  for $i = 1 : |\bar{D}|$ **do**
    $F_1, F_2 \leftarrow$ **Download**$(\theta_{\xi 1}^{t-1}, \theta_{\xi 2}^{t-1})$
    $(X, Y) \leftarrow D\ [i]$
    $Y_1, Y_2 \leftarrow F_1(X), F_2(X)$
    $(X, \hat{Y}_1, \hat{Y}_2) \leftarrow$ **Sample&Refine**$(X, Y_1, Y_2, Y, \epsilon)$
    $d_i \leftarrow (X, \hat{Y}_1, \hat{Y}_2)$
    $\Delta\theta_{i1}^t, \Delta\theta_{i2}^t, \mathcal{L}_i^t \leftarrow$ **Update**$(F_1, F_2; d_i)$
    $\overline{\theta}_{\xi 1}$.add$(\Delta\theta_{i1}^t)$, $\overline{\theta}_{\xi 2}$.add$(\Delta\theta_{i2}^t)$, $\overline{\mathcal{L}}$.add$(\mathcal{L}_i^t)$
  end
  $\theta_{\xi 1}^t, \theta_{\xi 2}^t \leftarrow$ **Aggregate**$(\overline{\theta}_{\xi 1}, \overline{\theta}_{\xi 2}, \overline{\mathcal{L}}; \beta, \lambda)$
end
return $\theta_{\xi 1}^T$ and $\theta_{\xi 2}^T$

C. Pseudo Label Generation and Selection

1) Pseudo Label Generation: Based on the cross-pseudo supervision [39], we train two differently initialized models, $F_1(.)$ and $F_2(.)$ to co-supervise each other with pseudo labels when pixel-level labels are not available. The training image X is fed to the two models F_1 and F_2 to generate pseudo labels Y_1 and Y_2 respectively. Y_1 and Y_2 are then refined, denoted as $\hat{Y}_1$ and $\hat{Y}_2$, and used for training each local client. Details of the corresponding refinement strategies for each type of label are introduced as follows:

1) **Pixel-level labels**: No pseudo labels, which can be expressed as $\hat{Y}_1 = \hat{Y}_2 = Y_{gt}$.
2) **Bounding box labels**: Each prediction pair $Y_1 = F_1(X)$ and $Y_2 = F_2(X)$ is refined according to the corresponding bounding box label, *i.e.*, $\hat{Y}_1 = Y_1 * Y_{bbox}$ and $\hat{Y}_2 = Y_2 * Y_{bbox}$.
3) **Image-level class labels**: Each prediction pair $Y_1 = F_1(X)$ and $Y_2 = F_2(X)$ is refined through the image-level label Y_{img}, *i.e.*, $\hat{Y}_1 = Y_1 * Y_{img}$ and $\hat{Y}_2 = Y_2 * Y_{img}$. In this way, images corresponding to only the background class are viewed as outliers and filtered out from being used for training.
4) **No labels** (*i.e.*, without supervision): Without available supervision, we let $\hat{Y}_1 = Y_1$, and $\hat{Y}_2 = Y_2$.

It should be noted that the main benefit of pseudo label refinement through image-level labels is the removal of background images. Given unlabeled data, if pseudo labels corresponding to both foreground and background images are treated equally for training without proper filtering/refinement, it will negatively affect representation learning to extract truly-useful information from unlabeled data. Removing those background

images according to image-level labels, *i.e.* outliers, is proven to be helpful.

A specific client i is trained by minimizing:

$$\mathcal{L}_i = \mathcal{L}_{dice}(Y_1, \hat{Y}_2) + \mathcal{L}_{dice}(Y_2, \hat{Y}_1), \tag{1}$$

where $\mathcal{L}_{dice}$ is the Dice loss function.

2) Dynamic Sample Selection: Despite the effectiveness of the above pseudo label generation and refinement processes, pseudo labels may still be incorrect. Training with incorrect pseudo labels can negatively impact the model performance [47]. Fortunately, prediction consistencies between the two federated models [73] have a positive correlation with the prediction's accuracy, which can be used for sample selection.

Given client i and its training data D_i, we aim to generate a mask $M_i = \{m_1, \ldots, m_{|D_i|} | m_i \in [0, 1]\}$ to select reliable training samples indicated by $m = 1$, according to

$$m_i = \begin{cases} 1 & \text{if } dice(\hat{Y}_1, \hat{Y}_2) \geq \epsilon \text{ and } Y \neq Y_{bbox} \\ 1 & \text{if } dice(\hat{Y}_1, \hat{Y}_2) \geq \epsilon \text{ and } dice(\hat{Y}_1, Y_{bbox}) \geq 0.5 \\ 1 & \text{if } \hat{Y}_1 = \hat{Y}_2 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\epsilon \in [0, 1]$ is a threshold that is inversely proportional to the number of selected training samples. For pixel-level supervision, $m_i = 1$ is set for all training samples as $\hat{Y}_1 = \hat{Y}_2 = Y_{gt}$. For bounding box supervision, Y_{bbox}, it is important to enforce $\hat{Y}_1$ to focus more on class-relevant information inside the bounding box regions, which is realized by imposing an additional constraint on $dice(\hat{Y}_1, Y_{bbox})$. As training progresses, the models are more capable of generating more accurate predictions. Then, $\sum_{i=1}^{i=|M_i|} m_i$ progressively increases to $|D_i|$, allowing the model to learn from a growing set of training data. More discussions of dynamic sample selection are provided in Section V-A.

D. Federated Model Update

At each federated training round t, each client first receives the federated model's parameters θ_ξ^{t-1} from the server. Each participating client then updates the model locally with its training data D_i. Finally, the gradient update from each local client $\Delta\theta_i^t$ is sent to the server to update the federated model's parameters by

$$\theta_\xi^t \leftarrow \theta_\xi^{t-1} + \sum_{i=1}^{|\bar{D}|} w_i \Delta\theta_i^t. \tag{3}$$

where w_i is the aggregation weight of each client, defined as $|D_i| / \sum_{i=1}^{i=|\overline{D}|} |D_i|$ in FedAvg [12] which ignores the inter-client variations in supervision strength. Since weakly-labeled and unlabeled data are more accessible than pixel-level labeled data, FedAvg assigns higher weights to unlabeled/weakly-labeled clients based on their data amounts. Consequently, the federated model focuses more on the less informative clients, which can be sub-optimal or even detrimental. One solution is to manually assign higher weights to the more informative clients [33], which is not trivial as the server may intrude into clients' privacy.

To effectively update the federated model with mixed supervision, in the early training rounds, it is critical to assign higher weights to the more informative clients, *e.g.*, those with more pixel-level labeled data. As training progresses, the federated model becomes more accurate and thus more robust to better utilize weakly-labeled and unlabeled data for training. Therefore, aggregation weights assigned to clients without pixel-level labeled data, *e.g.*, weakly-labeled and unlabeled clients, can be gradually increased. To this end, we utilize client-specific training loss for adaptive weight assignment. It is based on the observation that clients without pixel-level labeled data tend to have lower training loss as learning from pseudo labels generated using consistency regularization [39] is much easier than pixel-level supervised learning. Thus, adaptive weight assignment according to training loss not only prioritizes learning from more informative clients but also prevents the federated model from over-fitting. The proposed adaptive aggregation function is thus defined as

$$c_i \leftarrow \frac{|D_i|}{\sum_{i=1}^{i=|\overline{D}|} |D_i|}, d_i \leftarrow \frac{(\mathcal{L}_i)^\beta}{\sum_{i=1}^{i=|\overline{D}|} (\mathcal{L}_i)^\beta} \tag{4}$$

$$w_i \leftarrow \frac{c_i + \lambda \cdot d_i}{\sum_{i=1}^{i=|\overline{D}|} c_i + \lambda \cdot d_i}, \tag{5}$$

where λ and β are hyper-parameters to tune, impacting the degree of reliance on different clients. According to Eq. 5, the weight w_i assigned to a weakly-labeled or unlabeled client will gradually increase as the training progresses because: 1) the increasing amount of training samples used from this client (increasing the value of c_i) and 2) the reduction of training losses at pixel-level labeled clients (reducing the denominator of d_i). More discussions of adaptive aggregation can be found in Section V-A.

E. Revisit FedMix

To gain a better understanding of FedMix, we compare FedMix with known approaches from the following two perspectives:

1) **Federated Learning with Noisy Labels (FLNL).** FLNL does not rely on the assumption that each client has a clean dataset. To reduce the negative impact of label noise, existing FLNL approaches mainly employ label transition matrix estimation [70], regularization [71], loss function design [72], and clean sample selection [73], [74]. For this point, FedMix can be regarded as a special form of clean sample selection, where FedMix manages to select both clean samples and clean clients. There are also differences in problem assumptions: FLNL is often based on the assumption that noisy labels usually are manually generated while FedMix focuses on addressing the issue of noisy pseudo labels. In addition, clients in FLNL are assumed to have both clean and noisy labels while clients in FedMix with unlabeled/weakly-labeled data only have noisy labels.
2) **Federated Semi-supervised Learning (FSSL).** In FSSL, each client can have either fully-labeled or

TABLE I
STATISTICS OF THE BREAST ULTRASOUND DATASET

Site	# Patients	# Images	# Healthy / # Cancerous
BUS	600	780	133 / 647
BUSIS	562	562	0 / 562
UDIAT	163	163	0 / 163

TABLE II
STATISTICS OF THE HAM10K DATASET

Site	Source	# Patients	# Images
Rosendahl	rosendahl	1552	2259
Vidir	modern	1695	3363
	old	278	439
	molemax	3954	3954

unlabeled data. Through consistency training or self-supervised learning, hidden information among unlabeled data can be explored to assist fully-supervised learning. FedMix can be viewed as a generalized form of FSSL, namely prior knowledge guided FSSL, where the prior knowledge refers to the weak supervision for pseudo label refinement. Without weak supervision, FedMix would degenerate to FSSL.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

1) Dataset: Experiments are mainly carried out on two challenging medical image segmentation tasks:

- **Breast tumor segmentation.** In this task, three public breast ultrasound datasets, namely BUS [49], BUSIS [50], and UDIAT [51], are used for evaluation and each of them is regarded as a separate client. More details of this dataset are introduced in Table I.
- **Skin tumor segmentation.** HAM10K [52] consists of four different sources. Each source acts as a client in FL. The statistics of HAM10K are presented in Table II.

To study the mixed supervised federated setting, we generate weak labels, *e.g.*, bounding box labels and image-level labels, based on the pixel-level labels. Specifically, we generate bounding box supervision by first finding the smallest box that captures the tumor region and then following [69] to generate bounding boxes where a random margin of 1-10 pixels is added to each side of the bounding boxes. In this way, we can more realistically evaluate the robustness of FedMix to imprecise bounding box supervision.

Following the standard practice, five-fold cross validation of each source/client is adopted for evaluation. All the breast ultrasound and skin dermoscopy images are resized to 256×256 pixels and then randomly flipped and cropped to 224×224 pixels for training.

2) Evaluation Metrics: In this work, the Dice coefficient (DC) is used for the evaluation of the two segmentation tasks. Considering the two-model architecture of FedMix, the predictions/outputs of F_1 are used for evaluation.

B. Implementation Details

1) Network Architectures: UNet [53] combined with group norm [54] is selected as the baseline segmentation model.

2) Supervision Types: The following types of labels are included in our experiments: 1) pixel-level labels (denoted as L), 2) bounding box labels (denoted as B), 3) image-level class labels (denoted as I), and 4) unlabeled (denoted as U), *e.g.*, training with only the raw images.

3) Comparison Methods: The following four prevailing frameworks are included for comparison:

- Local learning (LL): Each client trains a private deep learning network based on its own data.
- Federated Averaging (FedAvg): All clients, owning pixel-level labels, collaboratively train a federated model.
- Semi-supervised federated learning via self-training [33] (FedST): FedST utilizes both pixel-level labeled and unlabeled data for federated training. FedST is chosen for comparison as it does not require additional information sharing beyond the locally-updated model parameters.
- Semi-supervised federated learning via gradient diversity reduction [35] (FedRGD): FedRGD minimizes gradient diversity across clients' models by replacing batch normalization with group normalization and employing a new model aggregation approach.
- Our proposed Federated learning with mixed supervision (FedMix): FedMix fully utilizes multiple types of supervision for federated training.

The performance of FedAvg under full supervision is regarded as an upper bound of federated learning techniques. We evaluate the performance of FedMix under the semi-supervised setting by comparing FedMix with FedST and FedRGD. We also evaluate the performance of FedMix under various settings to show how additional weak labels improve the federated model's performance.

4) Training Details: All the networks are initialized from scratch with Xavier initialization [40] and then trained using the Adam optimizer with a learning rate of 1e-3, a weight decay of 1e-4, and a batch size of 16. All methods are implemented within the PyTorch framework and trained on Nvidia GeForce Titan RTX GPUs for 300 federated training rounds. At each training round, each participating client uses the locally available data to perform local model updates for one training epoch, *i.e.*, one pass through its training data. Following standard practice, federated training is performed synchronously and the federated model parameters are updated every training round. We set $\epsilon = 0.9$, $\lambda = 10$, and $\beta = 1.5$ and $\beta = 2.5$ for adaptive aggregation on breast tumor and skin lesion segmentation respectively. The hyper-parameters are chosen after conducting grid search, *e.g.*, $\epsilon \in [0.8, 0.85, 0.9, 0.95]$, $\lambda \in [1, 5, 10]$, and $\beta \in [1.5, 2, 2.5]$.

C. Results on Breast Tumor Segmentation

1) Experiment Settings: Data from BUS, BUSIS, and UDIAT are represented by C1, C2, and C3 respectively. To better demonstrate the value of weak labels, C3, owning

TABLE III

QUANTITATIVE RESULTS OF DIFFERENT LEARNING FRAMEWORKS UNDER VARIOUS TYPES OF SUPERVISION SETTINGS FOR BREAST TUMOR SEGMENTATION. FEDADAPTAGG IS SHORT FOR THE FL FRAMEWORK BY REPLACING FEDAVG WITH THE PROPOSED ADAPTIVE AGGREGATION FUNCTION. TYPES OF CLIENTS CONSIDERED FOR EXPERIMENTS LISTED UNDER THE FIRST COLUMN, INCLUDING *U*: CLIENTS WITH ONLY UNLABELED DATA, *I*: CLIENTS USING IMAGE-LEVEL LABELS, *B*: CLIENTS USING BOUNDING BOX LABELS, AND *L*: CLIENTS USING PIXEL-LEVEL LABELS. "–" REPRESENTS THE BASELINE MODEL TO CALCULATE THE P-VALUE SCORES

Supervision [C1, C2, C3]	Method	DC (%)				p-value
		C1	C2	C3	Avg.	
$[L, L, L]$	LL	**78.3±2.8**	**91.6±0.7**	83.4±2.3	84.4±1.7	-
	FedAvg	77.3±2.3	91.3±0.7	85.6±1.3	84.7±0.7	<0.001
	FedAdaptAgg	77.9±1.7	90.9±1.0	**86.6±1.1**	**85.1±0.4**	<0.001
$[U, U, L]$	LL	65.9±3.1	85.4±1.1	83.4±2.3	78.3±1.4	-
	FedRGD	60.6±4.4	80.7±3.2	83.5±4.4	74.9±4.0	<0.001
	FedST	67.3±1.7	85.0±1.6	83.2±3.5	78.5±1.9	<0.001
	FedMix	**68.3±1.8**	**87.8±0.6**	**85.6±1.9**	**80.6±0.7**	<0.001
$[U, I, L]$	FedMix	69.3±2.5	87.8±0.8	85.3±2.0	80.8±0.8	-
$[I, U, L]$		68.9±1.8	88.6±0.4	85.9±1.9	81.1±0.2	0.009
$[I, I, L]$		70.0±2.0	88.7±0.8	85.8±1.4	81.3±0.6	0.003
$[U, B, L]$		68.0±0.3	89.4±0.4	88.1±0.5	81.8±0.3	0.008
$[B, U, L]$		**71.8±4.4**	88.7±0.7	85.3±0.6	81.9±1.7	<0.001
$[I, B, L]$		70.3±0.3	89.3±0.2	87.2±0.3	82.3±0.1	<0.001
$[B, I, L]$		69.6±0.4	89.1±0.3	87.7±0.5	82.1±0.4	<0.001
$[B, B, L]$		70.0±0.6	**89.9±0.4**	**89.6±0.1**	**83.2±0.3**	<0.001

the least amount of data, is selected as the client with pixel-level labels. The supervision types of C1 and C2 are adjusted accordingly for different cases. To further explore the benefits of incorporating mixed supervision under the federated setting, we split the healthy images from BUS, *i.e.* C1, into two sets, and assign one set to BUSIS, *i.e.* C2. In this way, both C1 and C2 can have image-level labels for better evaluation of FedMix.

2) Quantitative Evaluation: According to Table III, *e.g.*, with full supervision, the LL model of C2 has the highest DC score of 91.6%, indicating data homogeneity among its data. Comparatively, C1 and C3 perform slightly worse, *i.e.*, 78.3% and 83.4% respectively. While FedAvg is desirable compared to LL, their overall performance is quite close. With FedAdaptAgg, *i.e.* FedAvg combined with the proposed adaptive aggregation, for joint training, C3 benefits the most from the federation, achieving an increase of 2.2% in DC. Compared to LL and FedAvg, FedAdaptAgg achieves an average increase of 0.7% and 0.4% in DC in terms of overall segmentation performance.

Quantitative results of FedMix, FedRGD, and FedST under the semi-supervised setting, *i.e.* $[U, U, L]$, are provided in Table III. For LL, the results of C1 and C2 are produced by directly applying the model trained by C3. Compared to their locally-learned models with full supervision in Table III, C1 and C2 encounter severe performance degradation, *i.e.*, a reduction of 12.4% and 6.2% in DC respectively, mainly due to relatively limited training data of C3. One interesting observation is that the performance drop of C1 is much larger than that of C2, due to greater data heterogeneity between C1 and C3. The overall performance of FedRGD is even worse, especially for C1 and C2, as gradient diversity is much more severe for the image segmentation task. FedST and FedMix train better federated models compared to FedRGD, leading to an average increase of 3.6% and 5.7% in DC respectively. Compared to LL, though the overall performance of FedST is slightly better, both C2 and C3 encounter performance degradation, which would hinder the motivation of participation from C2 and C3. Comparatively, FedMix consistently improves the results of all three clients, achieving an average increase of 2.3% in DC compared to LL.

Quantitative results of FedMix under mixed supervision are summarized in Table III. Note that when C1 uses its image-level labels for training, not only C1 but also C2 and C3 would benefit from the federation, achieving an average increase of 0.8% and 0.3% in DC respectively. When both C1 and C2 use their image-level labels for training, the overall performance is further improved by 0.7% in DC compared to the semi-supervised setting. Progressively introducing stronger supervision to C1 and C2 improves the performance of FedMix. For instance, the average model performance is around 81.8% when one (either C1 or C2) has access to the bounding box labels and the other only has unlabeled data, *e.g.*, [U,B,L] and [B,U,L]. When introducing image-level labels to the corresponding unlabeled client, *e.g.*, [I,B,L] and [B,I,L], the performance of FedMix is further improved to 82.3% and 82.1% respectively. When C1 and C2 have access to bounding box labels, the overall performance of FedMix is quite close to those of LL and FedAvg based on full supervision. It should be noted that C2 and C3 benefit more from stronger supervision, due to greater data homogeneity between C2 and C3 which, in turn, makes FL more biased toward them than toward C1. As a result, the segmentation performance of C3 with weaker supervision is even better than that with full supervision.

To quantify the significance of performance improvements, we further calculate the p-value of each approach compared to the baseline under each supervision setting. Under both full and semi-supervision, $p < 0.001$ is satisfied for all approaches compared to LL. Under mixed supervision, introducing image-level supervision is relatively less significant than the introduction of bounding box supervision. It is quite reasonable as bounding boxes provide more information than image-level labels.

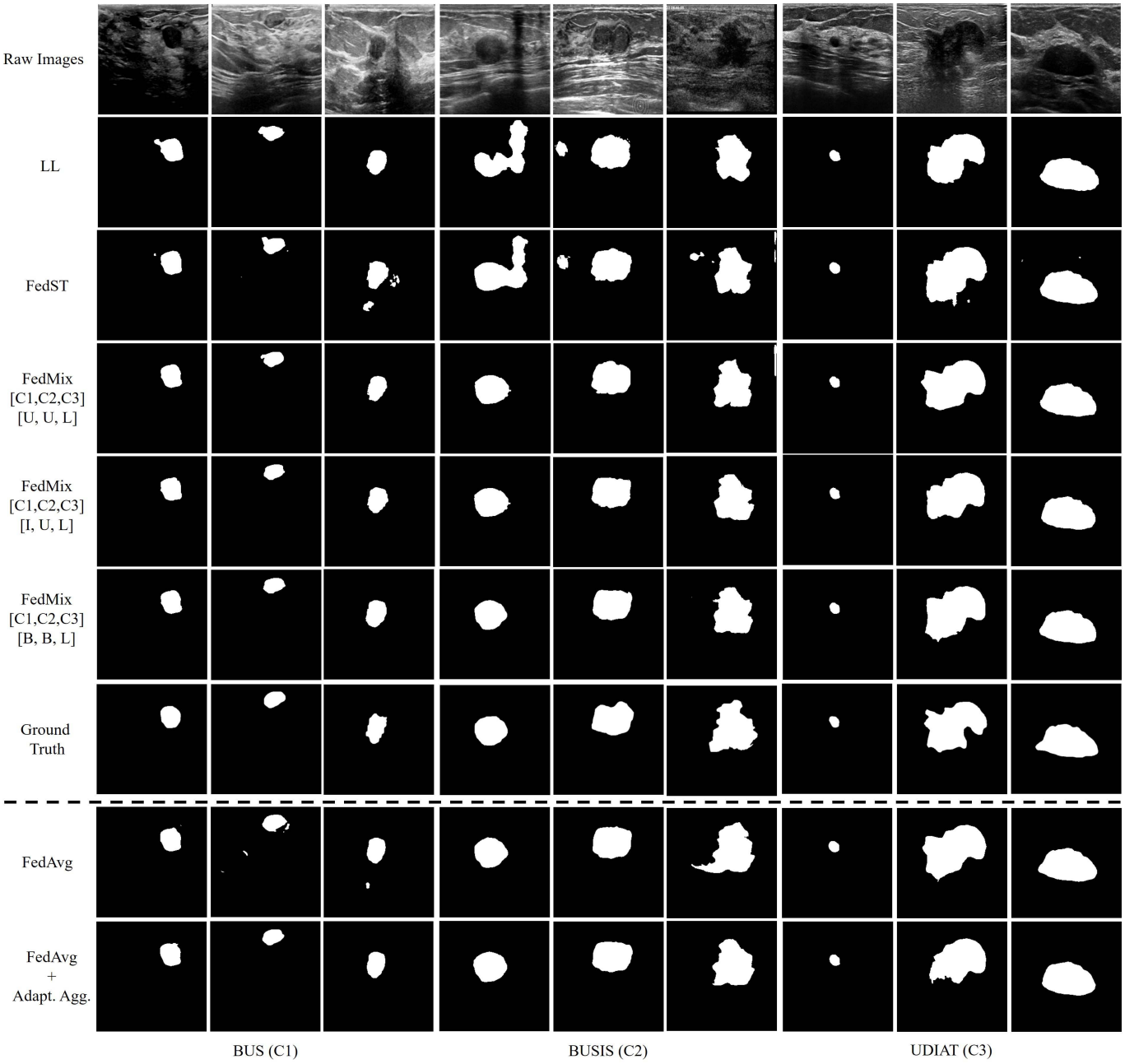

Fig. 3. Exemplar qualitative results of different learning frameworks for breast tumor segmentation. **The upper part** (Rows 1 to 7): the raw images, the segmentation maps produced by local learning (LL), FedST and FedMix under semi-supervision (*i.e.*,[C1, C2, C3] = [*U*, *U*, *L*]), the segmentation maps of FedMix under mixed supervision (*i.e.*,[C1, C2, C3] = [*I*, *U*, *L*] and [C1, C2, C3] = [*B*, *B*, *L*]), and the manual annotations by experts respectively. **The lower part** (Rows 8 to 9): the segmentation maps obtained by federated learning under full pixel-level supervision using FedAvg and the proposed adaptive aggregation function respectively.

3) Qualitative Evaluation: According to Fig. 3, LL trained on C3 produces quite a few false positives when testing on C2, indicating poor generalization capability due to limited training data. Under the semi-supervised setting, though the unlabeled data of C1 and C2 is used for training, the segmentation results of FedST are close to those of LL as learning from incorrect pseudo labels is not helpful and may be detrimental. Comparatively, FedMix can utilize the useful information in unlabeled data and the model generates predictions close to the experts' annotations. The introduction of stronger supervision signals (*i.e.*, from *U* to *I* and *B*) to FedMix would further reduce false positives and improve the shape preservation of tumor regions. Furthermore, adopting adaptive aggregation in federated learning is beneficial even under full supervision. The adaptively aggregated federated model can better capture the boundaries and shapes of the tumor regions and contain fewer false positives compared to the model learned by FedAvg.

D. Results on Skin Lesion Segmentation

1) Experiment Setting: Images from Rosendahl, Vidir-modern, Vidir-old, and Vidir-molemax are represented by C1,

TABLE IV

QUANTITATIVE RESULTS OF DIFFERENT LEARNING FRAMEWORKS UNDER VARIOUS TYPES OF SUPERVISION SETTINGS FOR SKIN LESION SEGMENTATION. FEDADAPTAGG IS SHORT FOR THE FL FRAMEWORK BY REPLACING FEDAVG WITH THE PROPOSED ADAPTIVE AGGREGATION FUNCTION. TYPES OF CLIENTS CONSIDERED FOR EXPERIMENTS LISTED UNDER THE FIRST COLUMN, INCLUDING *U*: CLIENTS WITH ONLY UNLABELED DATA, *I*: CLIENTS USING IMAGE-LEVEL LABELS, *B*: CLIENTS USING BOUNDING BOX LABELS, AND *L*: CLIENTS USING PIXEL-LEVEL LABELS. "–" REPRESENTS THE BASELINE MODEL TO CALCULATE THE P-VALUE SCORES

Supervision [C1, C2, C3, C4]	Method	DC (%)					p-value
		C1	C2	C3	C4	Avg.	
$[L, L, L, L]$	LL	88.7±0.4	92.9±0.4	93.0±1.2	94.8±0.3	92.6±0.3	-
	FedAvg	88.4±0.6	92.6±0.5	**95.7±0.8**	95.0±0.1	93.0±0.4	0.005
	FedAdaptAgg	**89.9±0.5**	**94.1±0.6**	95.2±0.8	**95.0±0.1**	**93.6±0.2**	<0.001
$[U, U, L, U]$	LL	74.9±1.0	73.0±1.2	93.0±1.2	91.1±1.7	83.3±0.7	-
	FedRGD	71.7±2.6	70.1±2.8	92.1±3.0	89.8±0.5	80.9±2.2	<0.001
	FedST	77.2±0.6	75.2±1.1	94.7±0.6	91.3±0.6	84.6±0.3	<0.001
	FedMix	**77.9±0.6**	**75.8±0.5**	**95.4±0.2**	**92.0±0.5**	**85.3±0.4**	<0.001
$[U, U, L, B]$	FedMix	77.9±2.2	80.0±2.8	**96.5±0.8**	91.5±0.7	86.5±1.3	-
$[U, B, L, B]$		82.9±1.4	88.4±1.6	96.2±0.2	91.3±1.4	89.7±1.1	<0.001
$[B, B, L, B]$		**85.7±0.2**	**89.4±0.7**	96.2±0.3	**93.5±0.3**	**91.2±0.1**	<0.001

C2, C3, and C4 respectively, and C3, owning the least amount of data, is selected as the client with pixel-level labels. The levels of labels on C1, C2, and C4 are adjusted accordingly under different settings.

2) Quantitative Results: From Table IV, under the fully-supervised setting, FedAvg improves the performance of the locally-learned models by an average of 0.4% in DC, indicating that cross-client collaboration is beneficial. Similar to breast tumor segmentation, we compare the proposed adaptive aggregation with FedAvg under full supervision for skin lesion segmentation. As stated in Table IV, adaptive aggregation effectively improves the performance of both C1 and C2 compared to FedAvg, leading to better overall segmentation performance, *i.e.* an increase of 0.6% in DC.

Under the semi-supervised setting where only C3 has access to pixel-level supervision (*i.e.*, L) as shown in Table IV, C3's locally-learned (LL) model does not perform well on C1 and C2, observed through the significant performance degradation which indicates severe inter-client variations between {C3, C4} and {C1, C2}. As a result, the pseudo labels on C1's and C2's data generated by the model trained by C3 may be inaccurate and utilizing them for training becomes harmful. Without effectively identifying those inaccurate pseudo labels, FedRGD obtains even worse segmentation performance across all the clients, resulting in an average decrease of 2.4% in DC compared to LL. Instead of using all the pseudo labels, FedST makes use of only confident predictions, leading to an average increase of 1.3% in DC compared to LL. With dynamic sample selection and adaptive aggregation, FedMix manages to select high-quality pseudo labels for model update and high-confident clients for model aggregation, thus improving the overall segmentation performance by an average of 2.0% and 0.7% in DC respectively compared to LL and FedST.

Quantitative results of FedMix under mixed supervision are presented in Table IV. Incorporating bounding box labels is helpful for pseudo label refinement. As a result, when more clients have access to bounding-box supervision, the segmentation performance of FedMix gradually improves, approaching the performance of FedAvg with full supervision. As bounding-box supervision is much more accessible than pixel-wise annotation, FedMix is more realistic in clinical scenarios.

In terms of p-value, under full supervision, performance improvements brought by data collaboration through FedAvg are less significant than those by the proposed FedAdaptAgg. Under both semi-supervision and mixed supervision, $p < 0.001$ is satisfied for all other approaches compared to the baseline. Quantitative p-value results demonstrate the stability of FedMix in dealing with various supervision settings.

3) Qualitative Results: Qualitative results of skin lesion segmentation are shown in Fig. 4. Consistent with the quantitative results, the segmentation maps on C1 and C2, produced by the locally-learned model on C3, are inaccurate, due to large inter-client variations between {C1, C2} and {C3, C4}. While the segmentation maps produced by FedST are slightly more accurate compared to LL, learning from confident pseudo labels is insufficient to train a generalizable model, as shown through the inaccurate segmentation maps produced by FedST on C1 and C2. Under the same supervision setting, FedMix produces more accurate segmentation maps by dynamically selecting high-quality pseudo labels for training. Given stronger supervision, *e.g.*, bounding box labels, FedMix improves the segmentation quality, especially on tumor shape preservation. Through the comparison under the fully-supervised setting, we observe that the segmentation maps produced by adaptive aggregation contain fewer false negatives and have better shape consistencies with manual annotations compared to FedAvg.

V. ABLATION STUDIES

A. Effectiveness of Each Component in FedMix

In this section, we evaluate the two key components in FedMix, namely adaptive aggregation and dynamic sample selection, through a series of ablation studies as summarized in Table V.

By setting λ to 0, FedMix relies only on dynamic sample selection, *i.e.* without adaptive aggregation. Consequently, the federated model may bias toward those noisy clients with inaccurate pseudo labels which is detrimental for convergence, making its segmentation performance even worse than that of directly applying the local model trained by one client to other clients, *i.e.* LL in Tables III and IV. On the other hand, by setting ϵ to 0, FedMix would not perform dynamic

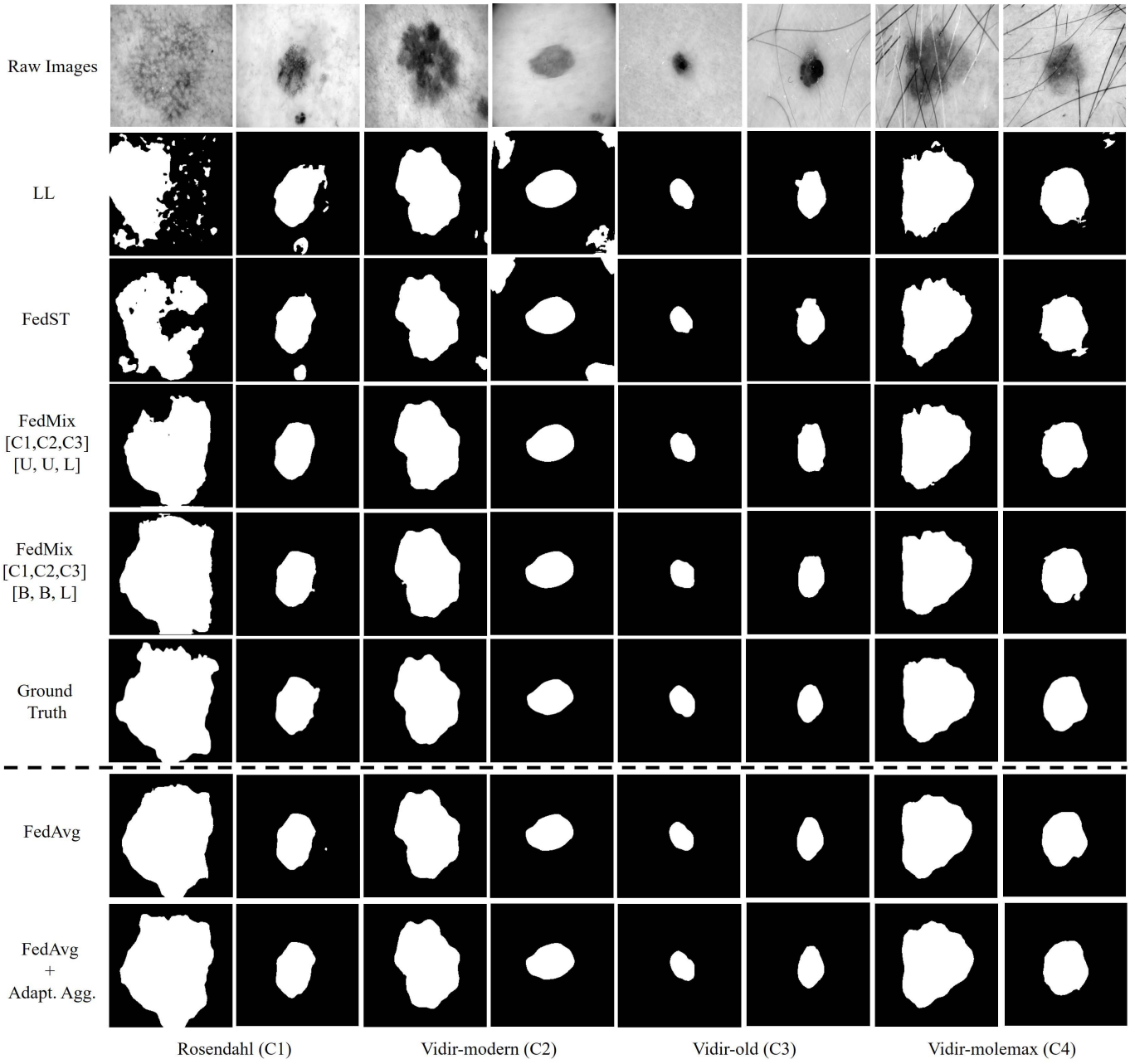

Fig. 4. Qualitative results of different learning frameworks for skin lesion segmentation. **The upper part** (Rows 1 to 6): the raw images, the segmentation maps produced by local learning (LL), FedST, FedMix under semi-supervision (*i.e.*,[C1, C2, C3, C4] = [*U*, *U*, *L*, *U*]), FedMix under mixed supervision (*i.e.*,[C1, C2, C3, C4] = [*B*, *B*, *L*, *B*]), and the expert annotations respectively. **The lower part** (Rows 7 to 8): the segmentation maps obtained by federated learning under the fully-supervised setting with FedAvg and the proposed adaptive aggregation function respectively.

sample selection. Compared to relying only on dynamic sample selection, adopting adaptive aggregation alone is much more helpful, as it will put more trust towards the clients with full supervision for the federated model update. In this way, interference from low-quality pseudo labels can be largely alleviated. When jointly using adaptive aggregation and dynamic sample selection, FedMix achieves the best segmentation performance, demonstrating the value of FedMix in properly leveraging training data from clients based on a wide range of supervision. We also calculate the p-values of different component combinations for further evaluation as listed in Table V. According to this analysis, both components in FedMix are proven valuable for consistent performance improvement.

One key hyper-parameter in dynamic sample selection is the threshold ϵ in dynamic sample selection, based on which a subset of training data would be selected for model update. To evaluate the impact of ϵ on overall segmentation performance, we conducted additional experiments by adjusting the value of ϵ as shown in Table VI. In general, with a smaller ϵ, more samples with relatively low confidence are included for model update, which in turn affects the training process

TABLE V
QUANTITATIVE RESULTS OF FEDMIX WITH THREE DIFFERENT COMBINATIONS OF USING THE TWO UNDERLYING COMPONENTS FOR BREAST TUMOR AND SKIN LESION SEGMENTATION UNDER SEMI-SUPERVISION. EXPERIMENTS WERE CONDUCTED USING 300 TRAINING ROUNDS FOR FIVE-FOLD CROSS VALIDATION OF EACH DATASET. "–" REPRESENTS THE BASELINE MODEL TO CALCULATE THE P-VALUE SCORES. *Adapt. Agg.* AND *Dyn. Sel.* ARE SHORT FOR ADAPTIVE AGGREGATION AND DYNAMIC SELECTION IN FEDMIX RESPECTIVELY

Components		Overall Performance			
Adapt. Agg.	Dyn. Sel.	Breast		Skin	
		DC	p-value	DC	p-value
×	✓	61.6±2.0	-	74.6±4.4	-
✓	×	78.6±1.5	<0.001	80.4±0.8	<0.001
✓	✓	**80.6±0.7**	<0.001	**85.3±0.4**	<0.001

TABLE VI
THE EFFECT OF THE THRESHOLD ϵ IN DYNAMIC SAMPLE SELECTION FOR BREAST TUMOR AND SKIN LESION SEGMENTATION (TRAINING WITH 100 ROUNDS FOR FIVE-FOLD CROSS VALIDATION OF EACH DATASET). *Adapt. Agg.* IS SHORT FOR ADAPTIVE AGGREGATION IN FEDMIX

Adapt. Agg.	ϵ	DC (%)	
		Breast	Skin
✓	0.9	**76.2±0.5**	**82.6±0.4**
	0.8	74.9±0.4	81.9±1.3
	0.7	74.4±0.5	79.7±0.7
	0.6	73.8±0.4	78.8±1.2
	0.5	74.6±0.5	78.5±0.8
	0.4	73.8±0.5	78.4±1.2
	0.3	75.2±0.7	79.0±0.6
	0.2	71.6±0.4	78.7±2.6
	0.1	72.2±0.5	78.8±1.2
	0.0	72.3±0.5	77.9±1.9

TABLE VII
THE EFFECT OF THE HYPER-PARAMETERS λ AND β IN ADAPTIVE AGGREGATION FOR BREAST TUMOR AND SKIN LESION SEGMENTATION (TRAINING WITH 300 ROUNDS FOR FIVE-FOLD CROSS VALIDATION OF EACH DATASET). SUPERVISION SETTINGS ARE [*U*, *U*, *L*] AND [*U*, *U*, *L*, *U*] FOR BREAST TUMOR AND SKIN LESION SEGMENTATION RESPECTIVELY. ϵ IS SET AS 0.9

λ	β	DC (%)	
		Breast	Skin
1	1.5	80.5±1.0	84.0±0.3
5	1.5	80.5±0.8	84.2±0.5
10	1.5	**80.6±0.7**	84.3±0.4
1	2.0	80.1±0.4	84.9±0.5
5	2.0	80.2±0.5	85.2±0.6
10	2.0	80.0±0.4	84.8±0.4
1	2.5	79.9±0.9	85.0±0.4
5	2.5	79.8±0.9	85.1±0.3
10	2.5	80.0±0.9	**85.3±0.4**

and results in a performance drop. With a larger ϵ, as only samples with relatively high confidence are selected for training, the overall segmentation performance is much improved. It demonstrates the value of dynamic sample selection in filtering out inaccurate pseudo labels for training.

Adaptive aggregation contains two hyper-parameters, *i.e.* λ and β, which determine how the training loss of each client affects its aggregation weight. As described above, we set the optimal values of λ and β through grid search as summarized in Table VII. In general, regardless of the settings, the segmentation performance on both breast tumor and skin lesion segmentation remains quite stable, varying within 0.8% and 1.1% respectively. This result validates that FedMix is not sensitive to the hyper-parameters. Between λ and β, the results show that λ is more deciding which is reasonable, as λ directly determines the weight of adaptive aggregation in federated model update.

TABLE VIII
STATISTICS OF THE RETOUCH [48] DATASET

Sources	# Volumes	# Images
Cirrus	24	3072
Topcon	22	2816
Spectralis	24	1176

B. Extension to Multi-Class Segmentation

To validate the extend-ability of FedMix to multi-class medical image segmentation, the publicly-available RETOUCH dataset, consisting of 112 OCT volumes with manual delineation of all three types of fluid, including intraretinal fluid (IRF), subretinal fluid (SRF), and pigment epithelial detachment (PED), is adopted for evaluation. Based on annotation availability, we conducted the experiments on the 72 OCT volumes as presented in Table VIII.

1) Experiment Setting: Slices from the three OCT devices in the RETOUCH dataset, namely Cirrus, Topcon, and Spectralis, are designated as C1, C2, and C3 respectively, and C3, possessing the least amount of data, is selected as the client with pixel-level labels. The levels of labels possessed by C1 and C2 are adjusted accordingly for different supervision settings. Here, we adopt three-fold cross validation for better comparison. For all experiments on the RETOUCH dataset, we initialize the networks from scratch with Xavier initialization [40] and then train them for 150 rounds using the Adam optimizer with a learning rate of 1.5e-3 and a weight decay of 1e-5. In terms of the hyper-parameters in FedMix, we set ϵ, λ, and β to 0.9, 2, and 3 respectively.

2) Quantitative Results: From Table IX, under full supervision, FedAvg outperforming LL demonstrates the necessity of data collaboration for better segmentation, especially when the data amount possessed by each client is limited. Compared to FedAvg, adopting the proposed adaptive aggregation for federated training can further improve the overall segmentation performance by 2.9% in DC.

Under the semi-supervised learning setting, relying only on the pixel-level labeled data of C3 is insufficient to extract domain-invariant features for multi-class segmentation, resulting in severe performance degradation compared to the results of LL with full supervision. Though FedRGD and FedST effectively improve the overall performance compared to LL, not all clients can benefit from FedST, notably C1 experiencing a reduction of 0.2% in DC. Comparatively, FedMix achieves consistent performance improvement across all clients, leading to the best overall performance under the semi-supervised setting.

TABLE IX

QUANTITATIVE RESULTS OF DIFFERENT FRAMEWORKS FOR MULTI-CLASS OCT FLUID SEGMENTATION UNDER VARIOUS SUPERVISION SETTINGS. TYPES OF CLIENTS CONSIDERED FOR EXPERIMENTS LISTED UNDER THE FIRST COLUMN, INCLUDING *U*: CLIENTS WITH ONLY UNLABELED DATA, *I*: CLIENTS USING IMAGE-LEVEL LABELS, *B*: CLIENTS USING BOUNDING BOX LABELS, AND *L*: CLIENTS USING PIXEL-LEVEL LABELS. "–" REPRESENTS THE BASELINE MODEL TO CALCULATE THE P-VALUE SCORES

Supervision [C1, C2, C3]	Method	DC (%)				p-value
		C1	C2	C3	Avg.	
$[L, L, L]$	LL	63.5±1.1	**59.5±0.4**	62.3±0.9	58.4±0.6	-
	FedAvg	66.4±1.0	56.1±1.8	76.5±2.2	66.3±1.2	<0.001
	FedAdaptAgg	**70.1±1.6**	58.7±1.3	**78.7±1.5**	**69.2±0.7**	<0.001
$[U, U, L]$	LL	52.7±1.1	38.8±1.2	62.3±0.9	51.3±0.5	-
	FedRGD	48.7±0.5	**43.2±2.0**	**68.1±0.9**	53.3±0.3	0.006
	FedST	52.5±3.0	41.7±1.4	65.4±1.3	53.2±1.9	<0.001
	FedMix	**54.3±1.7**	42.5±1.6	67.3±1.4	**54.7±0.8**	<0.001
$[I, I, L]$	FedMix	62.8±2.2	42.1±3.2	67.6±1.6	57.5±0.9	-
$[B, B, L]$		**65.8±3.5**	**45.9±1.3**	**71.2±0.9**	**61.0±1.3**	<0.001

With stronger supervision, FedMix achieves further performance improvement, especially for C1. With image-level labels, FedMix achieves an average increase of 2.8% in DC compared to semi-supervision. Incorporating bounding-box supervision produces further performance improvement for all three clients, approaching the performance of fully-supervised LL and FedAvg. The above quantitative results of multi-class segmentation being consistent with those of binary segmentation validate the flexibility and extend-ability of FedMix in dealing with mixed supervision for better medical image segmentation.

Similar to breast tumor and skin lesion segmentation, quantitative p-value results are calculated as summarized in Table IX. In general, $p < 0.001$ is satisfied in most cases, indicating convincing performance improvement brought by FedMix with both semi-supervision and mixed supervision.

C. Limitation and Future Work

Both dynamic sample selection and adaptive aggregation in FedMix require appropriate hyper-parameter tuning. For instance, though utilizing a stricter dynamic sample selection, *e.g.*, by setting $\epsilon = 0.9$, is highly beneficial as shown in Table VI, not using adaptive aggregation will degenerate the performance of FedMix to 61.6% and 74.6% in DC for breast tumor and skin lesion segmentation respectively, as shown in Table V. The optimal values of the hyper-parameters are task-dependent. Moreover, one of the implicit assumptions in adaptive aggregation is that clients with higher training losses are more likely to be more informative. However, clients with noisy/low-quality labels may also produce higher training losses, which can be hardly distinguished from the truly informative clients. Consequently, adaptive aggregation might become less effective in complicated scenarios. In the future work, we would explore additional information sharing for client re-weighting without privacy leakage.

Despite the success of FedMix in federated learning with mixed supervision, not all types of weak supervision, such as landmarks, scribbles, *etc.*, have been considered in this study. Based on the current design of FedMix, different types of supervision can be directly applied for pseudo label refinement. However, relying on pseudo labeling and refinement, inherent information in mixed supervision may not be fully utilized. In the future work, we would explore to adopt a unified multi-task learning architecture, *e.g.*. transformers [75], for various tasks/supervision.

VI. CONCLUSION

FedMix is the first federated learning framework that makes effective use of different levels of labels on each client for medical image segmentation. In FedMix, we first generate pseudo labels from clients and use supervision-specific refinement strategies to improve the accuracy and quality of pseudo labels. Then the high-quality data of each client is selected through dynamic sample selection for local model updates. To better update the federated model, FedMix utilizes an adaptive aggregation function to adjust the weights of clients according to both data quantity and data quality. Experimental results on two segmentation tasks demonstrate the effectiveness of FedMix on learning from various supervisions, which is valuable to reduce the annotation burden of medical experts. In the semi-supervised federated setting, FedMix outperforms the state-of-the-art approaches FedST and FedRGD. Compared to FedAvg, the proposed adaptive aggregation function achieves consistent performance improvements on the two tasks under the fully-supervised setting. More importantly, through validation, we demonstrate the extend-ability of FedMix for multi-class medical image segmentation. We believe the methods proposed in FedMix are widely-applicable in FL for medical image analysis beyond mixed supervision.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, "CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation," in *Proc. MICCAI*, 2018, pp. 732–740.

[3] Y. Li, L. Luo, H. Lin, H. Chen, and P.-A. Heng, "Dual-consistency semi-supervised learning with uncertainty quantification for COVID-19 lesion segmentation from CT images," in *Proc. MICCAI*, 2021, pp. 199–209.

[4] N. C. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging," in *Proc. ISBI*, 2018, pp. 168–172.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[6] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4981–4990.

[7] D. Zhang, G. Guo, W. Zeng, L. Li, and J. Han, "Generalized weakly supervised object localization," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 21, 2022, doi: 10.1109/TNNLS.2022.3204337.

[8] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1635–1643.

[9] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. ECCV*, 2016, pp. 549–565.

[10] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3159–3167.

[11] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. ICCV*, 2015, pp. 3213–3223.

[12] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016, *arXiv:1602.05629*.

[13] W. Li et al., "Privacy-preserving federated brain tumour segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2019, pp. 92–104.

[14] X. Li et al., "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med. Image Anal.*, vol. 65, Jan. 2020, Art. no. 101765.

[15] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12275–12284.

[16] J. Lee, J. Yi, C. Shin, and S. Yoon, "BBAM: Bounding box attribution map for weakly supervised semantic and instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2643–2652.

[17] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, and K.-T. Cheng, "Variation-aware federated learning with multi-source decentralized medical image data," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2615–2628, Jul. 2021.

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

[19] 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, pp. 1–11, 2021.

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

[21] Y.-X. Zhao, Y.-M. Zhang, M. Song, and C.-L. Liu, "Multi-view semi-supervised 3D whole brain segmentation with a self-ensemble network," in *Proc. MICCAI*, 2019, pp. 256–265.

[22] S. Reis, C. Seibold, A. Freytag, E. Rodner, and R. Stiefelhagen, "Every annotation counts: Multi-label deep supervision for medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9532–9542.

[23] P. Bilic et al., "The liver tumor segmentation benchmark (LITS)," in *Proc. MICCAI*, 2017, pp. 1–15.

[24] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Dec. 2014.

[25] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[26] J. Liu et al., "Active cell appearance model induced generative adversarial networks for annotation-efficient cell segmentation and identification on adaptive optics retinal images," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2820–2831, Jan. 2021.

[27] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, "Scribble-supervised video object segmentation," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 2, pp. 339–353, Feb. 2022.

[28] X. Li, L. Yu, H. Chen, C. Fu, and P. Heng, "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.

[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. (CVPR)*, Jun. 2020, pp. 12674–12684.

[30] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. H. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Proc. ECCV*, 2020, pp. 429–445.

[31] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proc. MICCAI*, 2020, pp. 552–561.

[32] D. Berthelot et al., "MixMatch: A holistic approach to semi-supervised learning," in *Proc. NeurIPS*, 2019, pp. 1–11.

[33] 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, May 2021, Art. no. 101992.

[34] Q. Liu, H. Yang, Q. Dou, and P.-A. Heng, "Federated semi-supervised medical image classification via inter-client relation matching," in *Proc. MICCAI*, 2021, pp. 325–335.

[35] Z. Zhang et al., "Improving semi-supervised federated learning by reducing the gradient diversity of models," in *Proc. ICBD*, 2021, pp. 1214–1225.

[36] T. Bdair, N. Navab, and S. Albarqouni, "FedPerl: Semi-supervised peer learning for skin lesion classification," in *Proc. MICCAI*, 2021, pp. 336–346.

[37] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu, "Federated contrastive learning for volumetric medical image segmentation," in *Proc. MICCAI*, 2021, pp. 367–377.

[38] N. Dong and I. Voiculescu, "Federated contrastive learning for decentralized unlabeled medical images," in *Proc. MICCAI*, 2021, pp. 378–387.

[39] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[41] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, "Cross-modality deep feature learning for brain tumor segmentation," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107562.

[42] D. Zhang et al., "Exploring task structure for brain tumor segmentation from multi-modality MR images," *IEEE Trans. Imag. Process.*, vol. 29, pp. 9032–9043, 2020.

[43] D. Zhang, W. Zeng, J. Yao, and J. Han, "Weakly supervised object detection using proposal- and semantic-level relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3349–3363, Jun. 2022.

[44] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5866–5885, Sep. 2022.

[45] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. NeurIPS*, 2020, pp. 1–13.

[46] W. Jeong, J. Yoon, E. Yang, and S. J. Hwang, "Federated semi-supervised learning with inter-client consistency & disjoint learning," in *Proc. ICLR*, 2021, pp. 1–15.

[47] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," in *Proc. ICLR*, 2021, pp. 1–20.

[48] H. Bogunovic et al., "RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 1858–1874, Feb. 2019.

[49] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.

[50] Y. Zhang et al., "BUSIS: A benchmark for breast ultrasound image segmentation," 2021, *arXiv:1801.03182*.

[51] M. H. Yap et al., "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218–1226, Aug. 2017.

[52] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, Aug. 2018.

[53] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[54] Y. Wu and K. He, "Group normalization," in *Proc. ECCV*, 2018, pp. 3–19.

[55] Y. B. Can et al., "Learning to segment medical images with scribble-supervision alone," in *Proc. MICCAI DLMIA*, 2018, pp. 236–244.

[56] M. Izadyyazdanabadi et al., "Weakly-supervised learning based-feature localization in confocal laser endomicroscopy glioma images," in *Proc. MICCAI*, 2018, pp. 300–308.

[57] J. Sun, A. Li, B. Wang, H. Yang, H. Li, and Y. Chen, "Soteria: Provable defense against privacy leakage in federated learning from representation perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9311–9319.

[58] 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, Jun. 2020.

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

[60] M. Y. Lu et al., "Federated learning for computational pathology on gigapixel whole slide images," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102298.

[61] I. Feki, S. Ammar, Y. Kessentini, and K. Muhammad, "Federated learning for COVID-19 screening from chest X-ray images," *Appl. Soft Comput.*, vol. 106, Jul. 2021, Art. no. 107330.

[62] N. Rieke et al., "The future of digital health with federated learning," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–7, 2020.

[63] 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. (CVPR)*, Jun. 2021, pp. 2423–2432.

[64] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "FedHealth: A federated transfer learning framework for wearable healthcare," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, Jul. 2020.

[65] C. Ju, D. Gao, R. Mane, B. Tan, Y. Liu, and C. Guan, "Federated transfer learning for EEG signal classification," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 3040–3045.

[66] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10713–10722.

[67] D. Li and J. Wang, "FedMD: Heterogenous federated learning via model distillation," 2019, *arXiv:1910.03581*.

[68] 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. (CVPR)*, Jun. 2021, pp. 1013–1023.

[69] H. Kervadec, J. Dolz, S. Wang, E. Granger, and I. B. Ayed, "Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision," in *Proc. MIDL*, 2020, pp. 365–381.

[70] J. Yao, H. Wu, Z. Zhang, I. W. Tsang, and J. Sun, "Safeguarded dynamic label regression for noisy supervision," in *Proc. AAAI*, 2019, pp. 9103–9110.

[71] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018, pp. 1–13.

[72] Y. Xu, P. Cao, Y. Kong, and Y. Wang, "L_DMI: A novel information-theoretic loss function for training deep nets robust to label noise," in *Proc. NeurIPS*, 2019, pp. 6222–6233.

[73] B. Han et al., "Coteaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. NeurIPS*, 2018, pp. 1–11.

[74] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13726–13735.

[75] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at SCA," 2021, *arXiv:2010.1192*.