International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-024-03139-6

ORIGINAL ARTICLE

Federated 3D multi-organ segmentation with partially labeled and unlabeled data

Zhou Zheng[1] · Yuichiro Hayashi[1] · Masahiro Oda[1,2] · Takayuki Kitasaka[3] · Kazunari Misawa[4] · Kensaku Mori[1,2,5]

Received: 24 March 2024 / Accepted: 28 March 2024

Abstract

Purpose This paper considers a new problem setting for multi-organ segmentation based on the following observations. In reality, (1) collecting a large-scale dataset from various institutes is usually impeded due to privacy issues; (2) many images are not labeled since the slice-by-slice annotation is costly; and (3) datasets may exhibit inconsistent, partial annotations across different institutes. Learning a federated model from these distributed, partially labeled, and unlabeled samples is an unexplored problem.

Methods To simulate this multi-organ segmentation problem, several distributed clients and a central server are maintained. The central server coordinates with clients to learn a global model using distributed private datasets, which comprise a small part of partially labeled images and a large part of unlabeled images. To address this problem, a practical framework that unifies partially supervised learning (PSL), semi-supervised learning (SSL), and federated learning (FL) paradigms with PSL, SSL, and FL modules is proposed. The PSL module manages to learn from partially labeled samples. The SSL module extracts valuable information from unlabeled data. Besides, the FL module aggregates local information from distributed clients to generate a global statistical model. With the collaboration of three modules, the presented scheme could take advantage of these distributed imperfect datasets to train a generalizable model.

Results The proposed method was extensively evaluated with multiple abdominal CT datasets, achieving an average result of 84.83% in Dice and 41.62 mm in 95HD for multi-organ (liver, spleen, and stomach) segmentation. Moreover, its efficacy in transfer learning further demonstrated its good generalization ability for downstream segmentation tasks.

Conclusion This study considers a novel problem of multi-organ segmentation, which aims to develop a generalizable model using distributed, partially labeled, and unlabeled CT images. A practical framework is presented, which, through extensive validation, has proved to be an effective solution, demonstrating strong potential in addressing this challenging problem.

Keywords Multi-organ segmentation · Federated learning · Semi-supervised · Partially supervised

Introduction

Accurate and robust multi-organ segmentation is highly required in computer-aided diagnosis, and its successive breakthroughs have been witnessed with the application of deep learning [1, 2]. To apply deep learning to multi-organ segmentation, one can collect a large-scale dataset with dense annotations from multiple institutes to train a generalizable model [3]. However, realizing such an application is usually restricted in clinical practice. On one aspect, medical datasets cannot be easily shared among medical institutes or

✉ Zhou Zheng
zzheng@mori.m.is.nagoya-u.ac.jp

✉ Kensaku Mori
kensaku@is.nagoya-u.ac.jp

[1] Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan

[2] Information Strategy Office, Information and Communications, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan

[3] School of Information Science, Aichi Institute of Technology, Yagusa-cho, 1247 Yachigusa, Toyota, Aichi, Japan

[4] Aichi Cancer Center Hospital, 1-1 Kanokoden, Chikusa-ku, Nagoya, Aichi, Japan

[5] Research Center for Medical Bigdata, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

Published online: 08 May 2024

clients due to privacy-keeping regulations. In addition, annotations for multi-organ can be incomplete and inconsistent across institutes. For instance, institutes may annotate single or partial organs that do not overlap with others due to different research interests. Another observation in practice is that institutes may leave many images unlabeled since annotating dense labels is costly. Thus, datasets for multi-organ segmentation are usually distributed since they cannot be shared and centralized. They are also imperfect since they do not have full annotations of multiple organs for fully supervised model training.

Driven by these observations, this work considers the problem of using distributed, partially labeled, and unlabeled samples to train a federated model for multi-organ segmentation. To the best of our knowledge, this problem remains unexplored. Three subproblems should be well addressed for this challenge: (1) learning from partially labeled samples, (2) learning from unlabeled samples, and (3) learning from distributed samples from multi-institute. To this end, this paper proposes FPS-Seg, a practical framework incorporating Federated learning (FL), Partially supervised learning (PSL), and Semi-supervised learning (SSL) modules for multi-organ Segmentation. Briefly, FPS-Seg maintains one central server and several clients. Clients locally train in-house models with partially labeled and unlabeled samples with PSL and SSL modules. The FL module bridges client and central server communication to prepare a global statistical model. With the collaboration of three modules, valuable information can be mined from imperfect local datasets and aggregated to develop a generalizable model. Contributions of this work are summarized in the following.

- A new problem setting, i.e., learning a model from decentralized, partially labeled, and unlabeled samples, is introduced for multi-organ segmentation, which is tougher and closer to clinical practice.
- A practical framework is designed to address this problem by unifying federated, partially supervised, and semi-supervised learning.
- The proposed method is extensively validated with several CT datasets. It shows a promising solution to this challenging problem. It also has a good generalization ability for downstream segmentation tasks.

Related works

Multi-organ segmentation remains a challenging task whose objective is to concurrently delineate multiple organs or anatomical structures from medical images, e.g., abdominal CT scans. Comprehensive insights into the domain can be gathered from dedicated reviews [1, 2].

The essence of semi-supervised learning (SSL) [4–6] is leveraging a small amount of labeled data alongside a much larger set of unlabeled data to train a model. Consistency learning [4, 5], expecting prediction invariance under perturbations, and pseudo-labeling [6], utilizing pseudo-labels for self-training, are two main strategies in SSL. Given the labor-intensive and costly nature of manual annotations in medical image analysis, SSL offers a viable alternative by tapping into the more accessible pool of unlabeled data. Several SSL methods [7, 8] have already been proposed for multi-organ segmentation.

Another key observation in practice is the substantial presence of datasets with only one or a few organs labeled in abdominal CT scans. To use these datasets with inconsistent and partially labeled annotations, a practical paradigm called partially supervised learning (PSL) has been introduced [9, 10]. While PSL is synonymous with SSL in some machine learning contexts [11], to clarify, this paper distinguishes between the two paradigms following prior works [9, 10]. In this work, SSL and PSL cater to different scenarios. SSL uses a mix of labeled and unlabeled data, whereas PSL manages datasets where each sample possesses some labels but not a full set.

Federated learning (FL) represents an advanced approach for decentralized data training, which is especially beneficial for sensitive fields like medical imaging, where datasets cannot be easily shared due to data privacy and regulations [12–14]. Several methods leveraging FL for multi-organ segmentation have been proposed. Notably, studies like [15, 16] have endeavored to train models on decentralized datasets with only partial annotations, combining PSL and FL for multi-organ segmentation.

Although great progress has been achieved by existing methods for multi-organ segmentation and other tasks in medical image analysis, these methods are primarily for a single task, i.e., SSL [7, 8], PSL [9, 10], and FL [13, 14], and dual tasks, e.g., federated semi-supervised learning [17, 18] and federated partial-label learning [15, 16]. Unlike previous works, this work introduces a more challenging and practical setting in multi-organ segmentation, which aims to learn a federated model from distributed, partially labeled, and unlabeled datasets by unifying SSL, PSL, and FL.

Method

Problem definition

Ideally, training a generalizable model for segmenting m organs requires numerous images $\mathbf{X}$ and the corresponding full annotations $\mathbf{Y}$ spanning $(m+1)$ classes, where $\mathcal{M} = \{0, 1, \ldots, m\}$ denotes the class set with $\{0\}$ for background and $\{1\}$ to $\{m\}$ for organs.

However, in clinical settings, datasets are often decentralized, with partial or no annotations. Given K medical institutes $\{Z_i\}_{i=1}^{K}$, each holds a dataset $\mathcal{D}_i = \left\{\mathcal{D}_i^u, \mathcal{D}_i^l\right\}$, where $\mathcal{D}_i^u = \left\{\mathbf{X}_i^u\right\}$ contains images $\mathbf{X}_i^u$ devoid of annotations and $\mathcal{D}_i^l = \left\{\mathbf{X}_i^l, \mathbf{Y}_i^l\right\}$ consists of images $\mathbf{X}_i^l$ and partial annotations $\mathbf{Y}_i^l$. This study considers an extreme case where each client only owns annotations for a single organ. Suppose that the label sets of $\left\{\mathbf{Y}_i^l\right\}_{i=1}^{K}$ are defined as $\{\mathcal{E}_i\}_{i=1}^{K}$, then $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \cdots \cap \mathcal{E}_K = \{0\}$, and $\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3 \cup \cdots \cup \mathcal{E}_K = \mathcal{M}$. These institutes are expected to utilize distributed, partially annotated, and unlabeled data to train a global model for multi-organ segmentation collaboratively.

Overview

The proposed framework FPS-Seg is shown in Fig. 1. FPS-Seg simulates a practice where a central server coordinates three medical institutes ($K = 3$) to collaboratively train a global model for multi-organ (liver, spleen, and stomach) segmentation. Institutes maintain teacher models $\{T_i\}_{i=1}^{K}$ and student models $\{S_i\}_{i=1}^{K}$. The teacher models use exponential moving averaging (EMA) weights of the student models. During local training phase, on one aspect, the student models learn from partially labeled samples. Besides, consistency is enforced between the outputs of teacher and student models to take advantage of unlabeled samples. The central server aggregates local student model weights to update the global model G. The FL, PSL, and SSL modules are introduced below.

Federated learning module

The FL module builds the bridge between local clients and the global server. Namely, it offers global model weight aggregation and local model weight updating functions. Its role is to train a global model $G(\cdot;\Theta^g)$ until convergence with a total of R rounds without data sharing to violate data privacy regulations. During training, at the r-th federated round, each institute of $\{Z_i\}_{i=1}^{K}$ would download the current global weight $\Theta_{(r)}^g$ from the server and assign it to the local model $S_i\left(\cdot;\Theta_i^s\right)$, which shares the same architecture as the global model. Afterward, clients fine tune local models for e epochs using their private datasets. The central server will then collect local weights $\left\{\Theta_{i(r)}^s\right\}_{i=1}^{K}$ and aggregates them to get updated global model weights $\Theta_{(r+1)}^g$. This study adopts federated average algorithm [19] to update the global model:

$$\Theta_{(r+1)}^g = \sum_{i=1}^{K} \frac{N_i}{\sum_{i=1}^{K} N_i} \Theta_{i(r)}^s, \tag{1}$$

where N_i denotes the number of images for each dataset $\mathcal{D}_i$ of client Z_i.

Partially supervised learning module

Assuming that the background, liver, spleen, and stomach class indexes are 0, 1, 2, and 3, the class set $\mathcal{M}$ is $\{0, 1, 2, 3\}$, and institutes Z_1, Z_2, and Z_3, respectively, hold private liver, spleen, and stomach datasets that comprise a large part of unlabeled samples and a small part of labeled samples. The label sets $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ are, respectively, $\{0, 1\}$, $\{0, 2\}$, and $\{0, 3\}$. The PSL module enables each client of $\{Z_i\}_{i=1}^{K}$ to train its local model $S_i\left(\cdot;\Theta_i^s\right)$ with partially labeled samples $\mathcal{D}_i^l = \left\{\mathbf{X}_i^l, \mathbf{Y}_i^l\right\}$.

Considering that a mini-batch of samples $\left\{\mathbf{x}_i^l, \mathbf{y}_i^l\right\}$ is fetched from $\mathcal{D}_i^l$, in which $\mathbf{x}_i^l \in \mathbb{R}^{B\times C\times H\times W\times D}$ denotes 3D CT volumes, where B, C, H, W, and D, respectively, indicate the size of the batch, channel, height, width, and depth, C is 1 for 3D CT volumes, and $\mathbf{y}_i^l \in \mathbb{R}^{B\times 2\times H\times W\times D}$ indicates corresponding partial annotations in one-hot formation for a specific organ. $S_i\left(\cdot;\Theta_i^s\right)$ outputs probability maps $\mathbf{p}_i^l \in \mathbb{R}^{B\times 4\times H\times W\times D}$ with the input of $\mathbf{x}_i^l$. The optimization objective for this module employs the marginal and exclusion losses as described in [10]. Please refer to [10] for more technical details. On one aspect, all unlabeled organs are treated as the background and merged into the original background, and a marginal loss $\mathcal{L}_{\text{marg}}$ is then calculated. In addition, the natural organ exclusiveness is added as additional prior knowledge to introduce a penalization in the form of an exclusion loss $\mathcal{L}_{\text{excl}}$.

The training procedure of client Z_1, as depicted in Fig. 2, is taken as an example, and note that other clients train models in a similar principle. Z_1 holds labeled samples $\mathcal{D}_1^l$ with annotations of the liver. The output probability maps of $S_1\left(\cdot;\Theta_1^s\right)$ are denoted as $\mathbf{p}_1^l$. Since the spleen and stomach are not labeled, their corresponding channels in $\mathbf{p}_1^l$ can be merged into the first channel, and the new probability maps $\hat{\mathbf{p}}_1^l \in \mathbb{R}^{B\times 2\times H\times W\times D}$ are then obtained. $\hat{\mathbf{p}}_1^l$ and $\mathbf{y}_1^l$ have the same channels, and a marginal loss $\mathcal{L}_{\text{marg}}$ can be calculated between them. Besides, the exclusive labels $\hat{\mathbf{y}}_1^l \in \mathbb{R}^{B\times 4\times H\times W\times D}$ are created for $\mathbf{p}_1^l$ based on $\mathbf{y}_1^l$. Expressly, for voxels belonging to the liver region in $\mathbf{x}_1^l$, the corresponding label values in $\hat{\mathbf{y}}_1^l$ are set to $[1, 0, 1, 1]$, while the remaining label values are set to $[0, 1, 0, 0]$. An exclusion loss $\mathcal{L}_{\text{excl}}$ is enforced between $\mathbf{p}_1^l$ and $\hat{\mathbf{y}}_1^l$ to reduce their intersection.

Generally, the training objective $\mathcal{L}_{\text{psl}}$ for each client of $\{Z_i\}_{i=1}^{K}$ is:

$$\mathcal{L}_{\text{psl}}\left(\mathbf{p}_i^l, \hat{\mathbf{p}}_i^l, \mathbf{y}_i^l, \hat{\mathbf{y}}_i^l\right) = \alpha\mathcal{L}_{\text{marg}}\left(\hat{\mathbf{p}}_i^l, \mathbf{y}_i^l\right) + \beta\mathcal{L}_{\text{excl}}\left(\mathbf{p}_i^l, \hat{\mathbf{y}}_i^l\right), \tag{2}$$

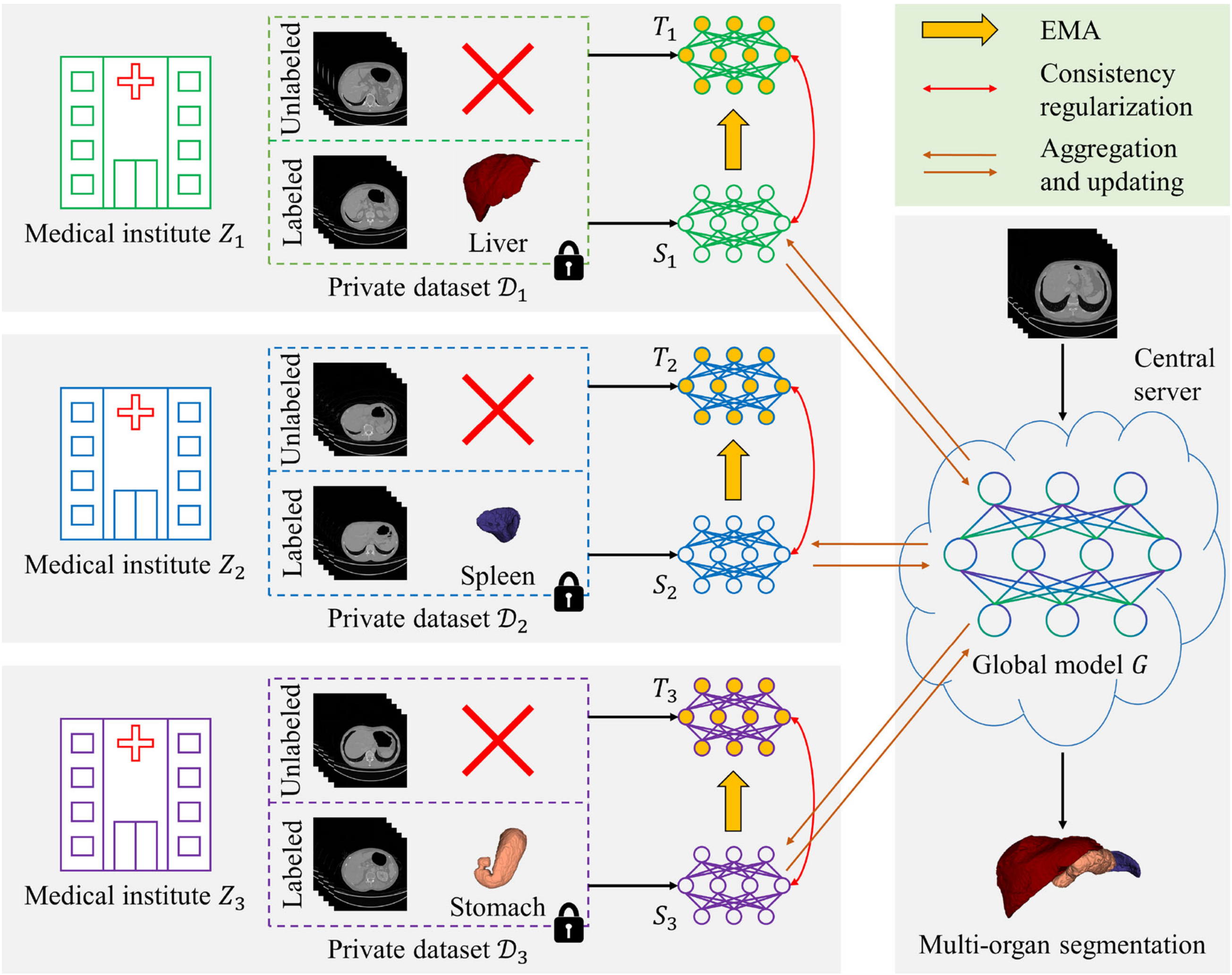


Fig. 1 Overview of proposed framework FPS-Seg

where

$$\mathcal{L}_{\text{marg}}\left(\hat{\mathbf{p}}_i^l, \mathbf{y}_i^l\right) = \underbrace{-\sum_{j=1}^{2}\sum_{v=1}^{V} \mathbf{y}_{i,j,v}^l \log \hat{\mathbf{p}}_{i,j,v}^l}_{\mathcal{L}_{\text{ce}}} \\ + \underbrace{\sum_{j=1}^{2}\left(1 - \frac{2\sum_{v=1}^{V} \hat{\mathbf{p}}_{i,j,v}^l \mathbf{y}_{i,j,v}^l}{\sum_{v=1}^{V}\left(\hat{\mathbf{p}}_{i,j,v}^l\right)^2 + \sum_{v=1}^{V}\left(\mathbf{y}_{i,j,v}^l\right)^2}\right)}_{\mathcal{L}_{\text{dice}}}, \quad (3)$$

and

$$\mathcal{L}_{\text{excl}}\left(\mathbf{p}_i^l, \hat{\mathbf{y}}_i^l\right) = \underbrace{\sum_{j=1}^{4}\sum_{v=1}^{V} \hat{\mathbf{y}}_{i,j,v}^l \log \mathbf{p}_{i,j,v}^l}_{\mathcal{L}_{\text{ece}}} \\ + \underbrace{\sum_{j=1}^{4} \frac{2\sum_{v=1}^{V} \mathbf{p}_{i,j,v}^l \hat{\mathbf{y}}_{i,j,v}^l}{\sum_{v=1}^{V}\left(\mathbf{p}_{i,j,v}^l\right)^2 + \sum_{v=1}^{V}\left(\hat{\mathbf{y}}_{i,j,v}^l\right)^2}}_{\mathcal{L}_{\text{edice}}}, \quad (4)$$

where j is the channel index, V is the number of voxels in an image, and v is the voxel index. α and β are hyperparameters. The combination of cross-entropy (CE) loss $\mathcal{L}_{\text{ce}}$ and Dice loss $\mathcal{L}_{\text{dice}}$ is adopted as the marginal loss $\mathcal{L}_{\text{marg}}$, and the combination of exclusion CE loss $\mathcal{L}_{\text{ece}}$ and exclusion Dice loss $\mathcal{L}_{\text{edice}}$ is employed as the exclusion loss $\mathcal{L}_{\text{excl}}$.

Semi-supervised learning module

The SSL module enables every client of $\{Z_i\}_{i=1}^K$ to further leverage its unlabeled samples $\mathcal{D}_i^u = \left\{\mathbf{X}_i^u\right\}$. Inspired by the work of [4], another model $T_i\left(\cdot; \Theta_i^t\right)$ is applied for each client of $\{Z_i\}_{i=1}^K$. $T_i\left(\cdot; \Theta_i^t\right)$ and $S_i\left(\cdot; \Theta_i^s\right)$ are regarded as

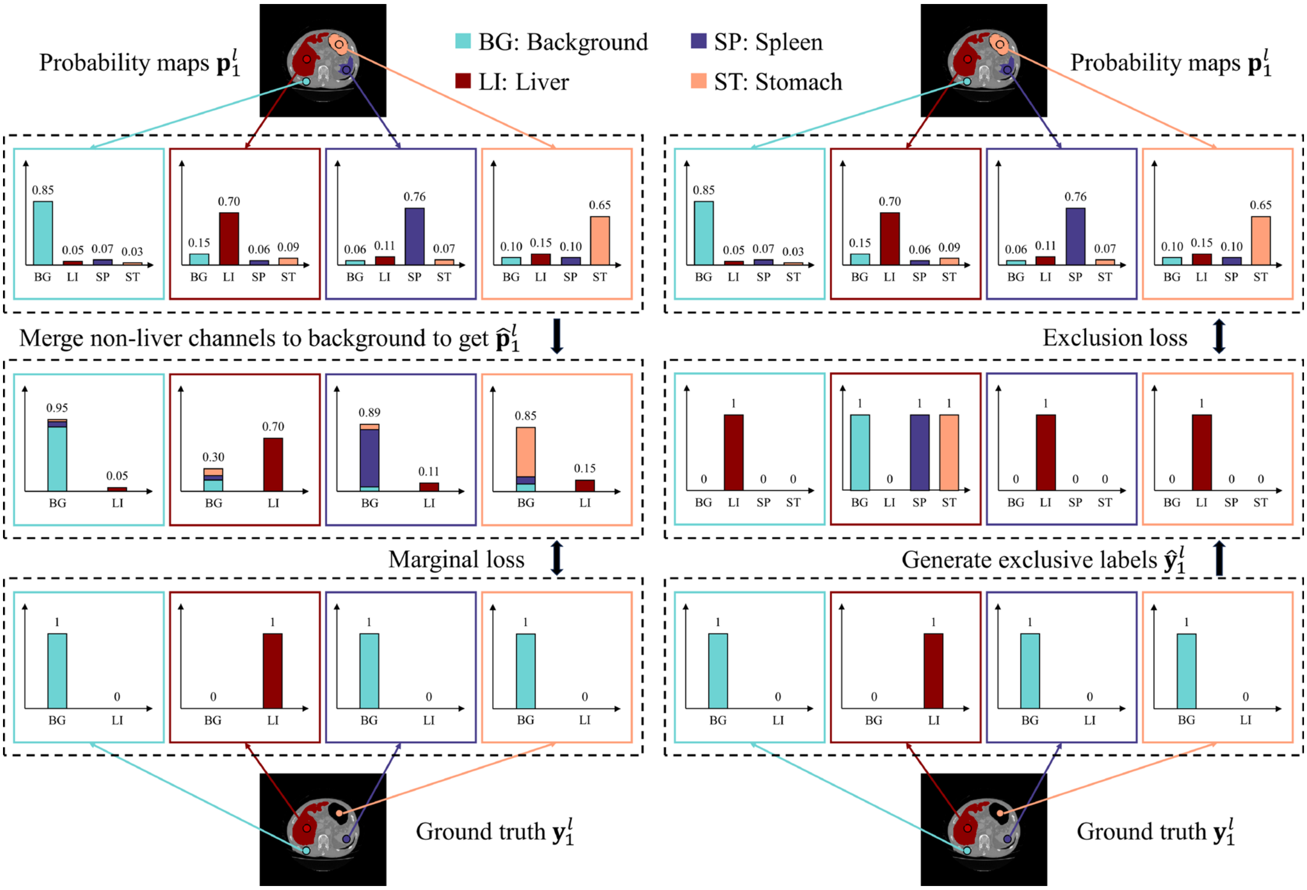

Probability maps $\mathbf{p}_1^l$
BG: Background
LI: Liver
SP: Spleen
ST: Stomach
Probability maps $\mathbf{p}_1^l$
0.85 0.05 0.07 0.03
0.15 0.70 0.06 0.09
0.06 0.11 0.76 0.07
0.10 0.15 0.10 0.65
BG LI SP ST
Merge non-liver channels to background to get $\widehat{\mathbf{p}}_1^l$
Exclusion loss
0.95 0.05
0.30 0.70
0.89 0.11
0.85 0.15
BG LI
Marginal loss
Generate exclusive labels $\widehat{\mathbf{y}}_1^l$
1 0
0 1
Ground truth $\mathbf{y}_1^l$
Ground truth $\mathbf{y}_1^l$

Fig. 2 Illustration of the PSL module, using the training procedure for client Z_1 as a representative example. For clarity, four voxels alongside hypothetical probability values are specified from the background, liver, stomach, and spleen to aid explanation

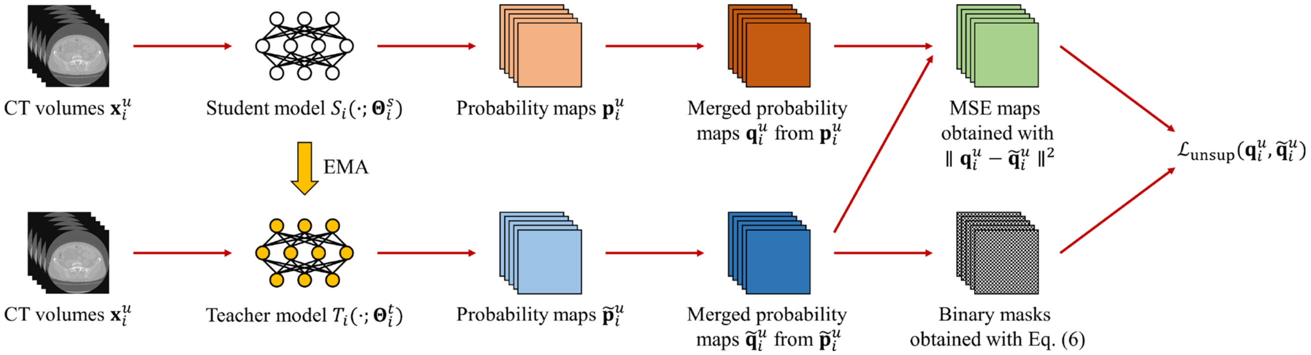

CT volumes $\mathbf{x}_i^u$
Student model $S_i(\cdot;\mathbf{\Theta}_i^s)$
Probability maps $\mathbf{p}_i^u$
Merged probability maps $\mathbf{q}_i^u$ from $\mathbf{p}_i^u$
MSE maps obtained with $\| \mathbf{q}_i^u - \tilde{\mathbf{q}}_i^u \|^2$
$\mathcal{L}_{\text{unsup}}(\mathbf{q}_i^u, \tilde{\mathbf{q}}_i^u)$
EMA
CT volumes $\mathbf{x}_i^u$
Teacher model $T_i(\cdot;\mathbf{\Theta}_i^t)$
Probability maps $\tilde{\mathbf{p}}_i^u$
Merged probability maps $\tilde{\mathbf{q}}_i^u$ from $\tilde{\mathbf{p}}_i^u$
Binary masks obtained with Eq. (6)

Fig. 3 Illustration of the training procedure for the SSL module

the teacher and the student models. The teacher model shares the same architecture as the student model and uses the student model's EMA weights. Consistency is imposed on their predictions for unlabeled data. Besides, input perturbation similar to the work [5] is further introduced since consistency regularization under harsher perturbations empirically benefits model generalization ability.

An illustration of the training procedure for the SSL module is shown in Fig. 3. Assuming that a mini-batch of unlabeled images $\mathbf{x}_i^u$ is fetched at each training iteration, these images are firstly fed into $T_i\left(\cdot;\Theta_i^t\right)$ and $S_i\left(\cdot;\Theta_i^s\right)$ to obtain probability maps $\tilde{\mathbf{p}}_i^u \in \mathbb{R}^{B\times 4\times H\times W\times D}$ and $\mathbf{p}_i^u \in \mathbb{R}^{B\times 4\times H\times W\times D}$. Same as Section “Partially supervised learning module,” the channels of unlabeled organs are then merged into the background for $\tilde{\mathbf{p}}_i^u$ and $\mathbf{p}_i^u$ to yield merged probability maps $\tilde{\mathbf{q}}_i^u \in \mathbb{R}^{B\times 2\times H\times W\times D}$ and $\mathbf{q}_i^u \in \mathbb{R}^{B\times 2\times H\times W\times D}$. Consistency learning regards $\tilde{\mathbf{q}}_i^u$ as pseudo-targets and calculates an unsupervised loss $\mathcal{L}_{\text{unsup}}$ between $\tilde{\mathbf{q}}_i^u$ and $\mathbf{q}_i^u$.

However, $\tilde{\mathbf{q}}_i^u$ may inevitably contain fault and noisy predictions, and consistency regulation based on which may accumulate training errors and result in model performance degradation. Confidence thresholding [5, 20], which involves setting a threshold τ, offers a practical solution to stabilize training and enhance model performance. It allows for the extraction of confident predictions from $\tilde{\mathbf{q}}_i^u$, enabling consistency regularization to rely solely on these predictions. By incorporating confidence thresholding, the training object on unlabeled samples $\mathcal{D}_i^u$ for each client of $\{Z_i\}_{i=1}^K$ is:

$$\mathcal{L}_{\text{unsup}}\left(\mathbf{q}_i^u, \tilde{\mathbf{q}}_i^u\right) = \frac{\sum_{j=1}^{2}\sum_{v=1}^{V}\Gamma_{i,v}\left\|\mathbf{q}_{i,j,v}^u - \tilde{\mathbf{q}}_{i,j,v}^u\right\|^2}{2\sum_{v=1}^{V}\Gamma_{i,v}}, \quad (5)$$

where

$$\Gamma_{i,v} = \begin{cases} 1 & \text{if } \max\left(\tilde{\mathbf{q}}_{i,1,v}^u, \tilde{\mathbf{q}}_{i,2,v}^u\right) > \tau \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

in which $\|\cdot\|^2$ is the mean error function (MSE) and $\Gamma_i \in \mathbb{R}^{B\times H\times W\times D}$ denotes the binary masks that control consistency regularization only using confident predictions. j is the channel index, V is the number of voxels in an image, and v is the voxel index. The threshold τ determines the extent of filtering, where $\tau = 0$ means all pseudo-target regions are included in the loss calculation, and $\tau = 1$ implies complete exclusion of pseudo-target regions.

Full training procedure of FPS-Seg

This part summarizes the full training procedure of FPS-Seg. At each federated round, each client first downloads the global model weight Θ^g and assigns it to student models $\left\{S_i\left(\cdot;\Theta_i^s\right)\right\}_{i=1}^K$. During the local training phase, student models $\left\{S_i\left(\cdot;\Theta_i^s\right)\right\}_{i=1}^K$ learn from labeled samples $\left\{\mathcal{D}_i^l\right\}_{i=1}^K$ with $\mathcal{L}_{\text{psl}}$, and extract information learns from unlabeled samples $\left\{\mathcal{D}_i^u\right\}_{i=1}^K$ with the help of $\left\{T_i\left(\cdot;\Theta_i^t\right)\right\}_{i=1}^K$ using $\mathcal{L}_{\text{unsup}}$. Thus, the total local training objective $\mathcal{L}_{\text{total}}$ for each client of $\{Z_i\}_{i=1}^K$ is:

$$\mathcal{L}_{\text{total}}\left(\mathbf{p}_i^l, \hat{\mathbf{p}}_i^l, \mathbf{y}_i^l, \hat{\mathbf{y}}_i^l, \mathbf{q}_i^u, \tilde{\mathbf{q}}_i^u\right) = \mathcal{L}_{\text{psl}}\left(\mathbf{p}_i^l, \hat{\mathbf{p}}_i^l, \mathbf{y}_i^l, \hat{\mathbf{y}}_i^l\right) + \gamma\mathcal{L}_{\text{unsup}}\left(\mathbf{q}_i^u, \tilde{\mathbf{q}}_i^u\right), \quad (7)$$

in which γ is a trade-off hyperparameter. When the local training finishes, the central server will aggregate local student weights $\left\{\Theta_i^s\right\}_{i=1}^K$ to update the global model $G(\cdot;\Theta^g)$ with Eq. (1). A global model can finally be obtained by repeating the above procedures.

Experiments and results

Experimental settings

Datasets and evaluation metrics

Datasets Three in-house contrast-enhanced abdominal CT datasets: #Set-A, #Set-B, and #Set-C, were applied. FPS-Seg was first evaluated with #Set-A, and its generalization ability was then validated by transferring it to downstream tasks on #Set-B and #Set-C. Details of three datasets are shown in Table 1. For data preprocessing, all volumes were resampled to an isotropic spatial resolution of 1.0 mm for each axis. The intensities were truncated to the range of [−1000, 1000] Hounsfield units (HU) and then normalized as zero mean and unit variance.

Evaluation metrics The Dice score [%] and 95% Hausdorff distance (95HD) [mm] were applied as evaluation metrics. The 95HD is a specific instance of the partial HD [21]. Given a surface point set $\mathcal{A}$ of the prediction and a surface point set $\mathcal{B}$ of the ground truth, the sets of directed HD from $\mathcal{A}$ to $\mathcal{B}$ and $\mathcal{B}$ to $\mathcal{A}$ are defined as

$$\omega(\mathcal{A},\mathcal{B}) = \left\{\min_{b\in\mathcal{B}}\|a-b\| \mid a\in\mathcal{A}\right\} \quad (8)$$

and

$$\omega(\mathcal{B},\mathcal{A}) = \left\{\min_{a\in\mathcal{A}}\|a-b\| \mid b\in\mathcal{B}\right\}, \quad (9)$$

respectively, where $\|\cdot\|$ denotes the Euclidean norm. The values $\omega_\kappa(\mathcal{A},\mathcal{B})$ and $\omega_\kappa(\mathcal{B},\mathcal{A})$ that rank in the κ-th percentile of $\omega(\mathcal{A},\mathcal{B})$ and $\omega(\mathcal{B},\mathcal{A})$ can then be chosen to calculate the partial HD $\Omega_\kappa(\mathcal{A},\mathcal{B})$ with

$$\Omega_\kappa(\mathcal{A},\mathcal{B}) = \max(\omega_\kappa(\mathcal{A},\mathcal{B}), \omega_\kappa(\mathcal{B},\mathcal{A})). \quad (10)$$

Table 1 Details of three datasets used in our study

Datasets	Organs	Volumes	Slices	Pixels	Resolutions
#Set-A	Liver, spleen, stomach	200	311 ∼ 1149	512 × 512	([0.59 ∼ 0.83] × [0.59 ∼ 0.83] × [0.50 ∼ 0.80]) mm^3
#Set-B	Pancreas	80	773 ∼ 1125	512 × 512	([0.63 ∼ 0.80] × [0.63 ∼ 0.80] × [0.50]) mm^3
#Set-C	Artery	80	468 ∼ 2532	512 × 512	([0.59 ∼ 0.98] × [0.59 ∼ 0.98] × [0.16 ∼ 1.00]) mm^3

Table 2 Quantitative results of the proposed method

Fourfold	Liver		Spleen		Stomach		Avg.	
	Dice	95HD	Dice	95HD	Dice	95HD	Dice	95HD
Fold-1	94.98	20.69	85.14	17.40	85.75	62.10	88.62 ± 5.52	33.40 ± 24.91
Fold-2	90.98	22.61	83.93	48.27	77.33	125.40	84.08 ± 6.83	65.43 ± 53.50
Fold-3	86.81	22.67	80.79	12.50	83.13	94.70	83.58 ± 3.04	43.29 ± 44.81
Fold-4	86.97	22.16	84.19	21.45	77.96	29.48	83.04 ± 4.61	24.37 ± 4.44
AVG.	89.94 ± 3.88	22.03 ± 0.92	83.51 ± 1.89	24.91 ± 16.00	81.04 ± 4.07	77.92 ± 41.36	84.83	41.62

Results are based on fourfold cross-validation

Table 3 Quantitative validation of various methods under localized, centralized, and federated learning settings, using different combinations of labeled (L) and unlabeled (U) samples

Method	Samples		Liver		Spleen		Stomach		Avg.	
	L	U	Dice	95HD	Dice	95HD	Dice	95HD	Dice	95HD
Localized Learning										
Client Z_1	50	0	92.16 ± 2.00	25.12 ± 9.23	–	–	–	–		
Client Z_2	50	0	–	–	79.19 ± 4.29	45.70 ± 17.12	–	–	84.92	37.50
Client Z_3	50	0	–	–	–	–	83.42 ± 6.80	41.67 ± 17.55		
Client Z_1	20	0	86.85 ± 4.13	28.90 ± 3.81	–	–	–	–		
Client Z_2	20	0	–	–	66.03 ± 18.34	52.18 ± 18.06	–	–	77.32	53.04
Client Z_3	20	0	–	–	–	–	79.09 ± 8.97	78.03 ± 32.60		
Client Z_1	20	30	88.36 ± 3.88	33.06 ± 3.72	–	–	–	–		
Client Z_2	20	30	–	–	75.09 ± 6.27	53.84 ± 29.11	–	–	81.76	46.55
Client Z_3	20	30	–	–	–	–	81.85 ± 7.22	52.76 ± 19.59		
Centralized learning										
FPS-Seg	50	0	92.19 ± 1.78	29.97 ± 12.10	87.08 ± 2.34	26.43 ± 10.95	83.54 ± 4.17	56.33 ± 34.62	87.61	37.57
FPS-Seg	20	0	89.15 ± 2.99	33.50 ± 8.50	81.90 ± 2.42	29.35 ± 16.32	77.64 ± 7.62	76.37 ± 61.18	82.89	46.41
FPS-Seg	20	30	90.49 ± 3.87	24.51 ± 3.88	82.72 ± 0.82	19.02 ± 5.27	79.89 ± 6.12	67.76 ± 54.67	84.36	37.10
Federated learning										
MTFL [16]	50	0	91.53 ± 2.76	13.26 ± 0.96	87.33 ± 2.40	11.42 ± 4.86	83.07 ± 4.93	40.74 ± 15.58	87.31	21.80
MTFL [16]	20	0	84.25 ± 6.93	25.09 ± 5.06	79.11 ± 4.91	23.82 ± 15.86	75.36 ± 6.78	57.40 ± 27.76	79.57	35.44
FPS-Seg	50	0	92.80 ± 2.09	16.56 ± 6.01	88.03 ± 2.37	20.08 ± 11.44	83.04 ± 5.29	44.21 ± 26.50	87.96	26.95
FPS-Seg	20	0	87.66 ± 4.45	25.82 ± 2.16	80.00 ± 4.21	30.97 ± 11.17	78.32 ± 5.04	91.83 ± 53.41	81.99	49.54
FPS-Seg	20	30	89.94 ± 3.88	22.03 ± 0.92	83.51 ± 1.89	24.91 ± 16.00	81.04 ± 4.07	77.92 ± 41.36	84.83	41.62

Table 4 Ablation study on FPS-Seg's components and corresponding hyperparameters

PSL	SSL	Confidence thresholding	α	β	γ	τ	Dice	95HD
✓	✗	✗	0	1	–	–	19.22	89.53
✓	✗	✗	1	0	–	–	77.42	64.22
✓	✗	✗	1	1	–	–	79.23	72.37
✓	✗	✗	1	2	–	–	79.86	73.44
✓	✗	✗	1	3	–	–	78.69	69.34
✓	✗	✗	1	4	–	–	78.10	99.77
✓	✗	✗	2	1	–	–	80.90	51.15
✓	✗	✗	3	1	–	–	80.94	53.03
✓	✗	✗	4	1	–	–	**81.99**	**49.54**
✓	✓	✗	4	1	0	–	81.99	49.54
✓	✓	✗	4	1	0.01	–	84.09	43.42
✓	✓	✗	4	1	0.05	–	84.14	43.68
✓	✓	✗	4	1	0.1	–	84.24	42.18
✓	✓	✗	4	1	0.5	–	84.61	42.13
✓	✓	✗	4	1	1	–	**84.83**	41.62
✓	✓	✗	4	1	5	–	84.80	**39.78**
✓	✓	✓	4	1	1	0	**84.83**	41.62
✓	✓	✓	4	1	1	0.8	84.78	**41.58**
✓	✓	✓	4	1	1	0.97	84.50	41.98

The optimal Dice score and 95HD value of each sub-step are highlighted in bold

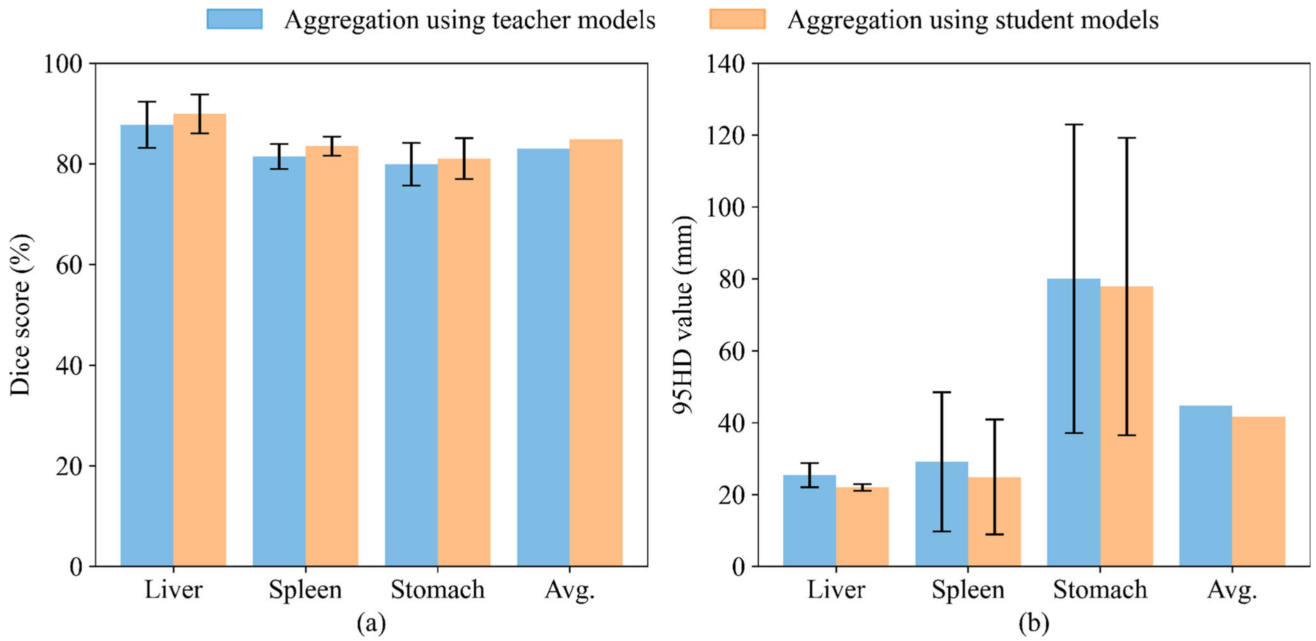


Fig. 4 Ablation study on aggregating student models and teacher models

κ is set as 95 to compute the 95HD.

Implementation details

Problem simulation One central server and three clients were maintained to simulate the problem. All experiments were conducted with fourfold cross-validation. #Set-A was split into 150/50 for training/validation at each fold. These 150 volumes were split into three sub-datasets (50/50/50) for three clients, and every sub-dataset was divided into 20 labeled samples and 30 unlabeled samples. Each sub-dataset only used annotations of a single organ.

Experimental setup All experiments were performed on the PyTorch platform. 3D U-Net [22] was chosen as the back-

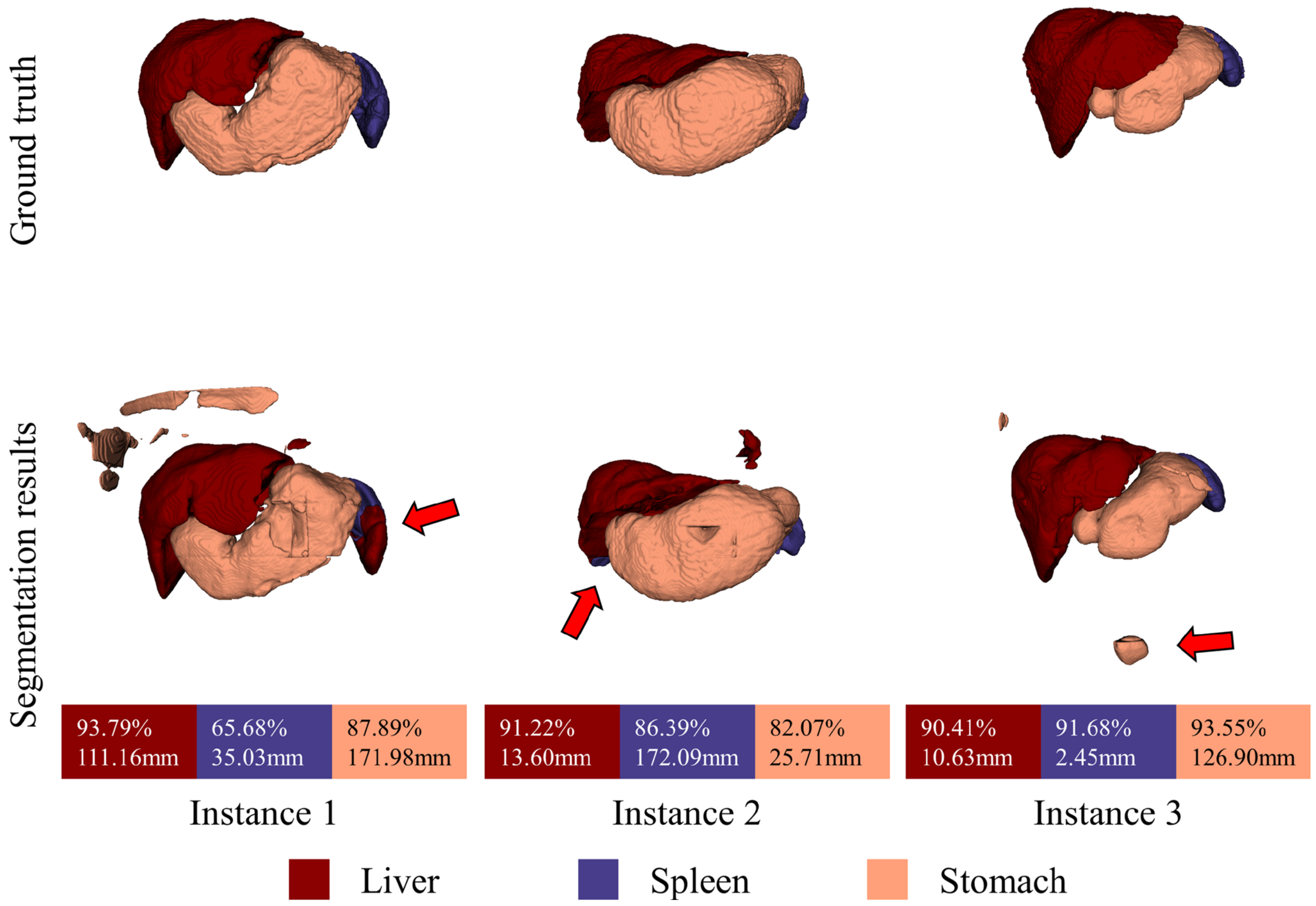


Fig. 5 Qualitative results of outliers. These instances are notable for achieving satisfactory Dice scores yet exhibiting large 95HD values for the liver, spleen, and stomach, respectively, as indicated by red arrows

bone. An SGD optimizer with a momentum of 0.9 and a weight decay of 10^{-4} was utilized to train the global model for 600 federated rounds. The local training epoch e was set to 1. A warm-up two-stage training strategy was adopted. Specifically, clients trained models using labeled samples under a poly-learning rate with an initial learning rate of 10^{-2} at the first 300 rounds and trained models using both labeled and unlabeled samples under a poly-learning rate with an initial learning rate of 10^{-3} at the second 300 rounds. Subvolumes with the size of $256 \times 256 \times 112$ were randomly cropped for training. Random flipping and random rotation were applied as augmentation schemes. For hyperparameter settings, please refer to Section "Ablation studies." During the testing phase, a sliding window strategy was applied.

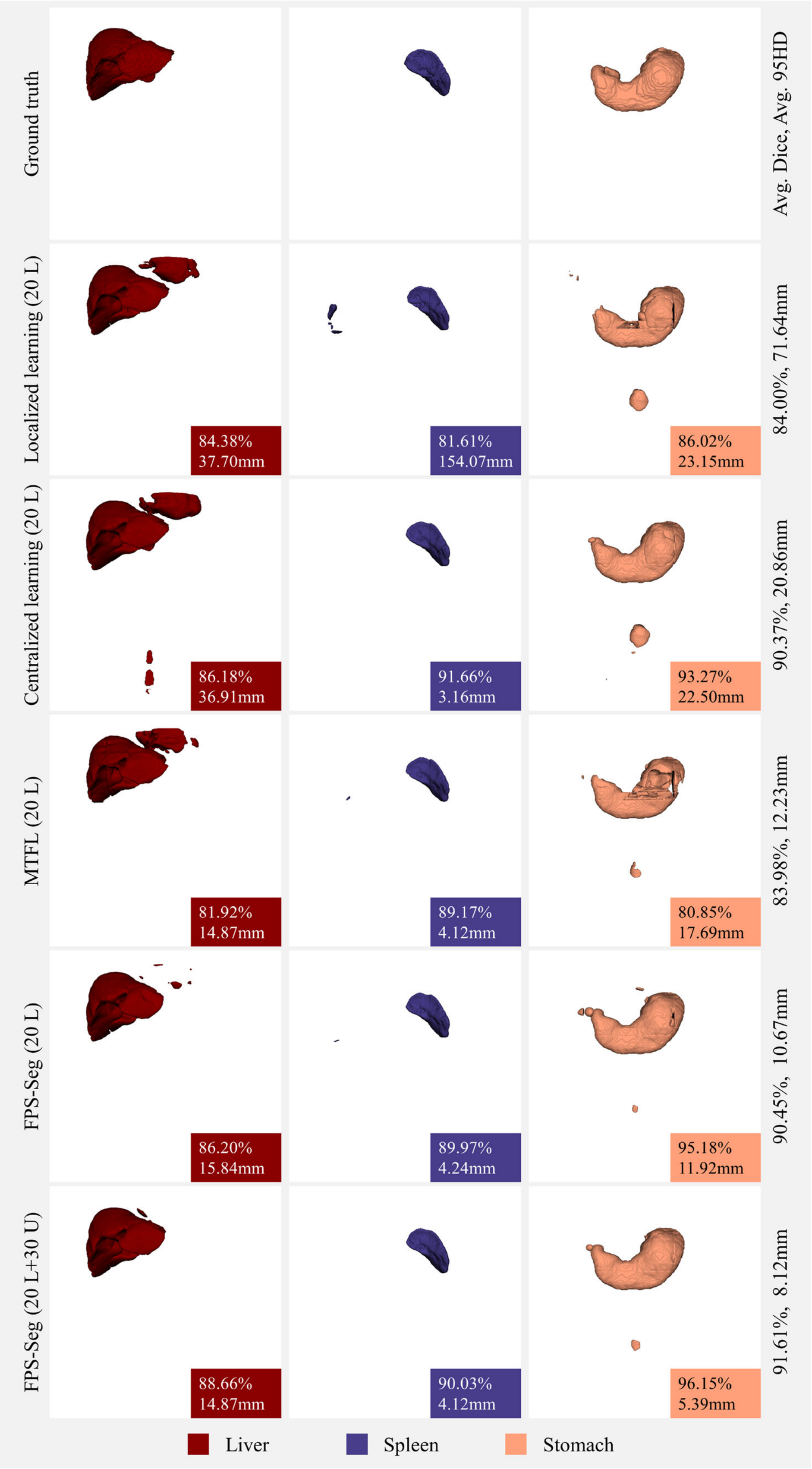
Experiment results

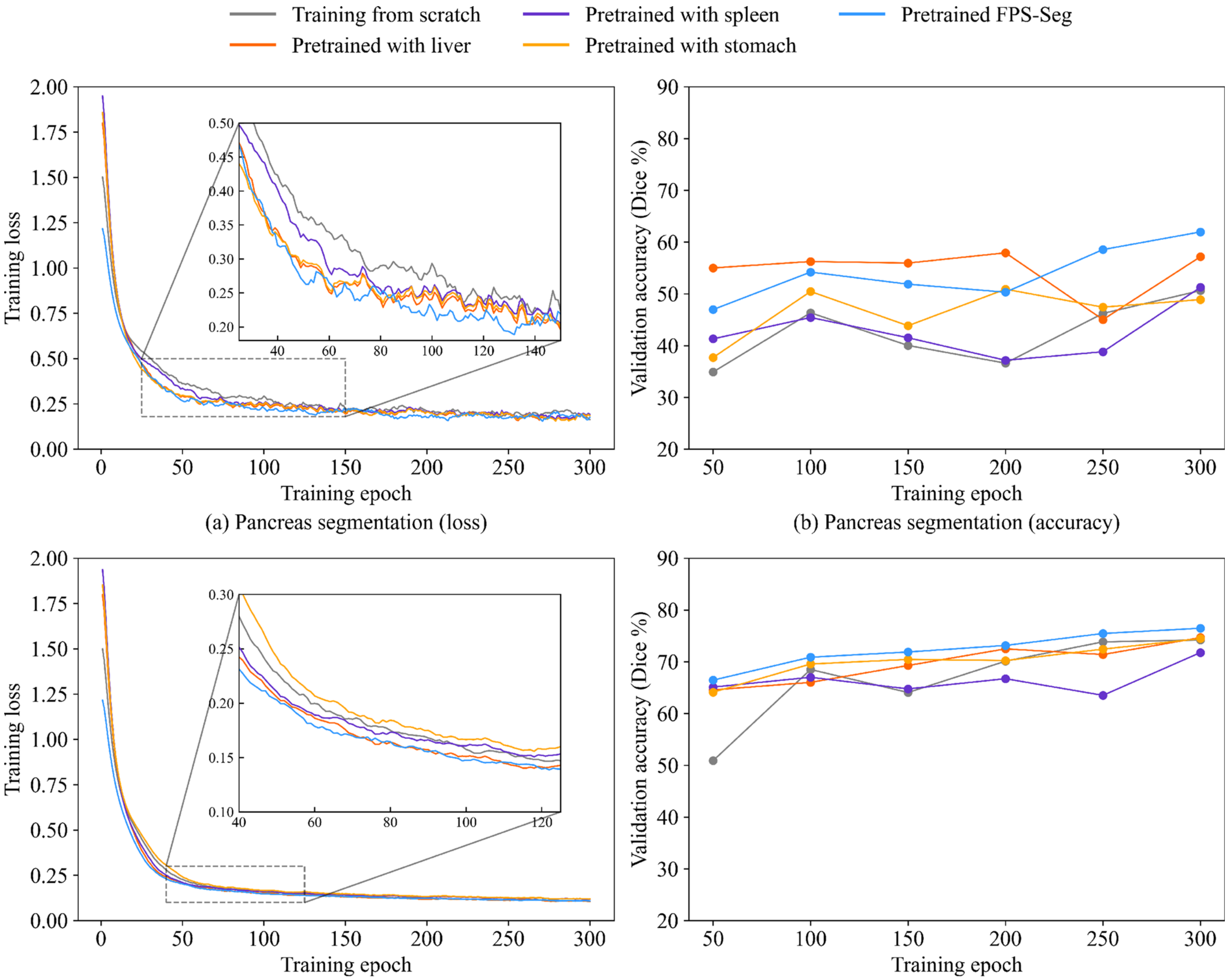
Quantitative results

Table 2 shows the quantitative results of FPS-Seg with four-fold cross-validation. Table 3 provides a detailed quantitative validation of different methods in localized, centralized, and federated learning scenarios. In the localized learning scenario, each client trained its local model on its private data with single organ annotations. Centralized learning involved training FPS-Seg using centralized datasets, employing PSL and SSL modules while excluding the FL module. The multitask federated learning (MTFL) approach [16] was implemented for comparison in the FL mode. These evaluations were conducted under three data scenarios: 50 L (50 labeled samples), 20 L (20 labeled samples), and 20 L + 30 U (20 labeled and 30 unlabeled samples).

Each method demonstrated its upper bound accuracy with 50 L. The performance of each method obtained with 20 L + 30 U consistently surpassed that with 20 L, validating the efficacy of SSL in utilizing unlabeled data. FPS-Seg consistently improved over localized learning, indicating its capability to exploit local datasets through FL. Additionally, FPS-Seg outperformed MTFL [16] in the FL mode and yielded competitive results comparable to its performance in the centralized learning mode.

Fig. 6 Qualitative results of various methods. 20 L: training with 20 labeled samples. 20 L + 30 U: training with 20 labeled and 30 unlabeled samples





(a) Pancreas segmentation (loss)

(b) Pancreas segmentation (accuracy)

(c) Artery segmentation (loss)

(d) Artery segmentation (accuracy)

Fig. 7 Comparison of convergence rate and validation accuracy in transfer learning for pancreas and artery segmentation across various methods

Ablation studies

Effects of FPS-Seg's components and their hyperparameters α and β in Eq. (2) are associated with the PSL module, γ in Eq. (7) relates to the SSL module, and τ in Eq. (6) is for confidence thresholding. This ablation study was divided into three sub-steps. Initially, the roles of α and β were examined with the PSL module using only labeled data. Once optimal values for α and β were established, the model incorporated unlabeled data by enabling the SSL module with varying γ. After determining suitable values for α, β, and γ, the model applied confidence thresholding with varying τ. This searching process allowed for an assessment of the individual contributions of each component, as well as evaluating the corresponding hyperparameters. Results are shown in Table 4. This paper set $\alpha = 4$, $\beta = 1$, $\gamma = 1$, and $\tau = 0$, under which FPS-Seg achieved superior performance.

Evaluating aggregation of student versus teacher models As depicted in Fig. 4, aggregating student models outperformed using teacher models. While teacher models maintain the EMA weights of the student models, this finding suggests that student models, which undergo direct gradient descent, are more effective for global model updating in our study.

Qualitative results

Three distinct instances with outliers are visualized in Fig. 5. These instances are notable for achieving satisfactory Dice scores yet exhibiting large 95HD values for the liver, spleen, and stomach, respectively, as indicated by red arrows. This visualization underscores the challenge FPS-Seg faces in certain instances where segmentation results contain outliers.

The qualitative results of various methods are displayed in Fig. 6. With 20 L, FPS-Seg outperformed localized training, surpassed the FL method MTFL [16], and showed results on par with centralized learning. Moreover, incorporating 30 U into the training further enhanced FPS-Seg's performance.

Transfer to downstream tasks

Initially pretrained on #Set-A, FPS-Seg was transferred to pancreas and artery segmentation on #Set-B and #Set-C, respectively. The datasets were divided into 60/20 for training/validation for both #Set-B and #Set-C. Comparisons were conducted among 3D U-Net trained from scratch, initialized with weights obtained by pretraining on single organs such as the liver, spleen, and stomach, and initialized with pretrained FPS-Seg. These comparisons were drawn throughout 300 epochs until convergence was reached. As depicted in Fig. 7, models initialized with pretrained FPS-Seg exhibited faster convergence and superior validation performance compared to those trained from scratch and those pretrained on single organs across the two downstream tasks.

Discussion and conclusion

This paper introduced a challenging multi-organ segmentation problem, which was considered based on the following observations in reality: (1) datasets cannot be easily shared, and thus, we cannot collect a large-scale dataset to train a generalizable model; (2) a large part of images is unlabeled across institutes since annotation is costly; and (3) only a small number of images may be partially labeled, and annotations are inconsistent across institutes due to different research targets. Training a generalizable model using these distributed, partially labeled, and unlabeled samples is highly required in clinical practice and remains unexplored.

A practical approach, FPS-Seg, was introduced to tackle this problem. FPS-Seg comprised three key modules: partially supervised learning, semi-supervised learning, and federated learning modules. These modules respectively, managed to learn from partially labeled, unlabeled, and distributed samples. This method was straightforward in addressing partially supervised, semi-supervised, and federated learning in a unified way. Extensive experiments were conducted to show FPS-Seg's successful solution for this challenging problem and good generalization ability for downstream segmentation tasks.

The proposed method was evaluated with liver, spleen, and stomach segmentation in CT images. Extending this method to segment additional organs using various modalities is considered an avenue for future work.

Acknowledgements This work was supported by the JSPS KAKENHI Grant Nos. 21K19898 and 17H00867, the JST CREST Grant No. JPMJCR20D5, and the JSPS Bilateral International Collaboration Grants, Japan.

Funding Open Access funding provided by Nagoya University.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This study was approved by the institutional review boards of the Nagoya University and the Aichi Cancer Center Hospital.

References

1. Cerrolaza JJ, Picazo ML, Humbert L, Sato Y, Rueckert D, Ángel González Ballester M, Linguraru MG (2019) Computational anatomy for multi-organ analysis in medical imaging: a review. Med Image Anal 56:44–67
2. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X (2021) A review of deep learning based methods for medical image multi-organ segmentation. Phys Med 85:107–122
3. Ji Y, Bai H, GE C, Yang J, Zhu Y, Zhang R, Li Z, Zhanng L, Ma W, Wan X, Luo P (2022) AMOS: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Advances in neural information processing systems
4. Tarvainen A, Valpola H (2017) Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems, vol 30
5. French G, Laine S, Aila T, Mackiewicz M, Finlayson G (2020) Semi-supervised semantic segmentation needs strong, varied perturbations. In: British machine vision conference
6. Zou Y, Zhang Z, Zhang H, Li C-L, Bian X, Huang J-B, Pfister T (2021) Pseudoseg: designing pseudo labels for semantic segmentation. In: International conference on learning representations
7. Zhou Y, Wang Y, Tang P, Bai S, Shen W, Fishman E, Yuille A (2019) Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In: 2019 IEEE winter conference on applications of computer vision. IEEE, pp 121–140
8. Xia Y, Yang D, Yu Z, Liu F, Cai J, Yu L, Zhu Z, Xu D, Yuille A, Roth H (2020) Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med Image Anal 65:101766
9. Zhou Y, Li Z, Bai S, Wang C, Chen X, Han M, Fishman E, Yuille AL (2019) Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10672–10681
10. Shi G, Xiao L, Chen Y, Zhou SK (2021) Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Med Image Anal 70:101979

11. Liu B (2007) Partially supervised learning. In: Web data mining: exploring hyperlinks, contents, and usage data. Springer, Berlin, pp 151–182
12. Li T, Sahu AK, Talwalkar A, Smith V (2020) Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 37(3):50–60
13. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, Ourselin S, Sheller M, Summers RM, Trask A, Xu D, Baust M, Cardoso MJ (2020) The future of digital health with federated learning. NPJ Digit Med 3(1):1–7
14. Kaissis GA, Makowski MR, Rückert D, Braren RF (2020) Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell 2(6):305–311
15. Xu X, Deng HH, Gateno J, Yan P (2023) Federated multi-organ segmentation with inconsistent labels. IEEE Trans Med Imaging 42(10):2948–2960
16. Shen C, Wang P, Yang D, Xu D, Oda M, Chen P-T, Liu K-L, Liao W-C, Fuh C-S, Mori K, Wang W, Roth HR (2022) Joint multi organ and tumor segmentation from partial labels using federated learning. In: Distributed, collaborative, and federated learning, and affordable AI and healthcare for resource diverse global health, pp 58–67
17. Yang D, Xu Z, Li W, Myronenko A, Roth HR, Harmon S, Xu S, Turkbey B, Turkbey E, Wang X, Zhu W, Carrafiello G, Patella F, Cariati M, Obinata H, Mori H, Tamura K, An P, Wood BJ, Xu D (2021) Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Med Image Anal 70:101992
18. Kassem H, Alapatt D, Mascagni P, Karargyris A, Padoy N (2022) Federated cycling (FedCy): semi-supervised federated learning of surgical phases. IEEE Trans Med Imaging
19. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics, pp 1273–1282
20. French G, Mackiewicz M, Fisher M (2018) Self-ensembling for visual domain adaptation. In: International conference on learning representations
21. Huttenlocher DP, Klanderman GA, Rucklidge WJ (1993) Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 15(9):850–863
22. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical image computing and computer-assisted intervention, LNCS, vol 9901, pp 424–432

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.