Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Medical Image Analysis 70 (2021) 101992

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Challenge Report

Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan☆

Dong Yang[a,1], Ziyue Xu[a,1], Wenqi Li[a], Andriy Myronenko[a], Holger R. Roth[a], Stephanie Harmon[b,c], Sheng Xu[d], Baris Turkbey[b], Evrim Turkbey[e], Xiaosong Wang[a], Wentao Zhu[a], Gianpaolo Carrafiello[f], Francesca Patella[g], Maurizio Cariati[g], Hirofumi Obinata[h], Hitoshi Mori[h], Kaku Tamura[h], Peng An[i], Bradford J. Wood[d], Daguang Xu[a,*]

[a] Nvidia Corporation, 4500 East West Highway, Bethesda, Maryland 20814, USA
[b] Molecular Imaging Branch, National Cancer Institute, NIH, Bethesda, MD, USA
[c] Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Molecular Imaging Branch, National Cancer Institute, NIH, Bethesda, MD USA
[d] Center for Interventional Oncology, Radiology and Imaging Sciences, NIH Clinical Center and National Cancer Institute, Center for Cancer Research, National Institutes of Health, Bethesda, MD, USA
[e] Radiology and Imaging Sciences, NIH Clinical Center, National Institutes of Health, Bethesda, MD, USA
[f] Radiology Department, Fondazione IRCCS Cá Granda Ospedale Maggiore Policlinico, University of Milan, Italy
[g] Diagnostic and Interventional Radiology Service, San Paolo Hospital; ASST Santi Paolo e Carlo, Milan, Italy
[h] Self-Defense Forces Central Hospital, Tokyo, Japan
[i] Department of Radiology, Xiangyang First People's Hospital Affiliated to Hubei University of Medicine Xiangyang, Hubei, China

ARTICLE INFO

Article history:
Received 2 June 2020
Revised 18 December 2020
Accepted 1 February 2021
Available online 6 February 2021

Keywords:
COVID-19
Chest CT
Federated learning
Semi-supervision

ABSTRACT

The recent outbreak of Coronavirus Disease 2019 (COVID-19) has led to urgent needs for reliable diagnosis and management of SARS-CoV-2 infection. The current guideline is using RT-PCR for testing. As a complimentary tool with diagnostic imaging, chest Computed Tomography (CT) has been shown to be able to reveal visual patterns characteristic for COVID-19, which has definite value at several stages during the disease course. To facilitate CT analysis, recent efforts have focused on computer-aided characterization and diagnosis with chest CT scan, which has shown promising results. However, domain shift of data across clinical data centers poses a serious challenge when deploying learning-based models. A common way to alleviate this issue is to fine-tune the model locally with the target domains local data and annotations. Unfortunately, the availability and quality of local annotations usually varies due to heterogeneity in equipment and distribution of medical resources across the globe. This impact may be pronounced in the detection of COVID-19, since the relevant patterns vary in size, shape, and texture. In this work, we attempt to find a solution for this challenge via federated and semi-supervised learning. A multi-national database consisting of 1704 scans from three countries is adopted to study the performance gap, when training a model with one dataset and applying it to another. Expert radiologists manually delineated 945 scans for COVID-19 findings. In handling the variability in both the data and annotations, a novel federated semi-supervised learning technique is proposed to fully utilize all available data (with or without annotations). Federated learning avoids the need for sensitive data-sharing, which makes it favorable for institutions and nations with strict regulatory policy on data privacy. Moreover, semi-supervision potentially reduces the annotation burden under a distributed setting. The proposed framework is shown to be effective compared to fully supervised scenarios with conventional data sharing instead of model weight sharing.

☆ NIH has a cooperative research and development agreement with NVIDIA that involves artificial intelligence and deep learning using medical imaging. This research was supported in part by the Center for Interventional Oncology, NIH Grant#1ZIDBC011242 &1ZIACL040015 and the Intramural Research Program of the NIH.

* Corresponding author.
E-mail address: daguangx@nvidia.com (D. Xu).

https://doi.org/10.1016/j.media.2021.101992
1361-8415/

1. Introduction

The COVID-19 pandemic has caused millions of confirmed cases and hundreds of thousands of deaths globally. Early and effective diagnosis is among critical measurements to control the infectious disease and manage its spread. Current standard test for potential SARS-CoV-2 infection is via RT-PCR. However, this technique can have high false negative/positive rate for coronavirus due to multifactorial issues (Drosten et al., 2003).

As a complimentary approach, medical imaging, specifically chest CT, is undergoing investigations on its capability of revealing important characteristics of COVID-19. Although its specificity for diagnostic purposes lacks consensus and is not recommended according to the guidelines from the ACR American College of Radiology, researchers have been looking for visual patterns related to viral infection from the images. A better understanding of the appearance of COVID-19 in chest CT may serve as an epidemiologic tool for mitigating outbreaks. Current consensus on the imaging pattern is the presence of infiltrates, from ground glass opacity to consolidation, usually bilateral, multilobar, with a peripheral distribution (Guan et al., 2020; Bernheim et al., 2020).

Quantification and characterization over the region of infiltrates may further help the understanding and tracking of disease progression, and provide important information of this novel virus in an outbreak situation (Colombi et al., 2020). In order to achieve this goal, detection and delineation of disease patterns are required. Unfortunately, such processes can be extremely tedious and time-consuming, especially at this time when many experts are already working around the clock on their clinical duty during a pandemic. Therefore, automated computerized methods are being developed and deployed to facilitate the understanding of the findings from medical images (Shi et al., 2020). Among them, methods based on deep learning techniques often achieve the state-of-the-art accuracy (Li et al., 2020a; Fan et al., 2020; Ouyang et al., 2020; Wang et al., 2020; Kang et al., 2020).

To design and develop an artificial intelligence (AI) system that can properly and robustly handle this problem, large amounts of data, as well as annotations, from diverse sources are required. However, in reality: 1) data access is often limited by strict sharing policies on sensitive private patient information, and 2) annotations have great variances in both quantity and quality due to experts' experience, availability and cost across imaging sites. Hence, most existing methods are trained using limited amount of data from a single site. On the other hand for model deployment, a well-known challenge for deep learning is the "domain shift" caused by the distribution difference between source data and target data, often leading to significant performance variance or degradation among different sites. Therefore, a mechanism enabling cross-institution collaboration under the constraint of significant difference in annotation availability and strict data sharing policies is highly desirable and may facilitate AI model development for COVID-19.

In this work, we propose a novel system to address the aforementioned challenges, which is based on federated and semi-supervised learning. Federated learning (Li et al., 2020c), especially ones with secure features (Li et al., 2019), can often give sufficient flexibility to different institutions to collaboratively train deep learning models without data sharing, as shown in Fig. 1; while semi-supervised learning can ensure effective training even when some sites have only limited amount of annotated data but large amount of unannotated data. In addition, the semi-supervised setting could reduce the burden of experts annotation, which is very valuable in current pandemic situation.

To test the proposed framework, we chose the task of segmentation for abnormal regions related to COVID-19, which is the most time-consuming as compared with other tasks like classification. This is because for most cases, slice-by-slice delineation is needed. With various configurations of the proposed framework, we show that our method's design can naturally benefit from multiple heterogeneous data sources under semi-supervise scenario. Finally, our method is task and training pipeline independent which makes it easy to be adapted to other deep learning tasks, such as COVID classification in CXR/CT, and non-COVID imaging tasks.

2. Related work

Federated Learning is an advanced distributed learning concept that takes advantage of datasets across multiple institutions without any explicit training data centralization or sharing (Yang et al., 2019; Li et al., 2020c). Although federated learning (FL) was initially designed for mobile edge devices, it has attracted increasing attention in healthcare domain because of its privacy preserving nature of the patient information. FL is agnostic to the type of the input data. It is capable of analyzing various medical data modalities, from free-text clinical reports to high-dimensional medical images (Xu and Wang, 2019). Brisimi et al. (2018) adopted FL to train a predictive model and solve a support vector machine problem for analysis of electronic health record (EHR) data. FL was applied for wearable healthcare using personalized machine learning models (Chen et al., 2020). Li et al. (2020c) built an FL framework for multi-site fMRI classification with preserved privacy. Recently, FL has been successfully applied on multi-institutional brain MRI for tumour segmentation with deep neural networks and improved privacy preserving of patient information (Sheller et al., 2018; Li et al., 2019).

Semi-Supervised Learning leverages the available information of unlabeled data, together with the supervision from labeled data, to improve the effectiveness and generalizability of machine learning models. In computer vision, semi-supervised learning has been investigated from different perspectives for various applications (e.g. image recognition). To take advantage of unlabeled data, consistency constraints have been investigated to mitigate the gap within and between domains of labeled and unlabeled data (Verma et al., 2019; Berthelot et al., 2019b; 2019a; Sohn et al., 2020; Liu et al., 2020a). One research trend is the framework of teacher-student models, which perform well utilizing consistency constraints between models for labeled and unlabeled data, respectively (Tarvainen and Valpola, 2017; Luo et al., 2018). Another similar framework, "noisy-student", achieved the state-of-the-art performance on ImageNet classification when jointly trained with a large amount of unlabeled data (Xie et al., 2019). Meanwhile, consistency-based model regularization can be implemented using model predictions of unlabeled data with data augmentation or pre-processing (Berthelot et al., 2019b; 2019a; Verma et al., 2019; Sohn et al., 2020). Another trend is to design auxiliary supervised tasks for unlabeled data, such as solving jigsaw puzzles (Noroozi and Favaro, 2016), predicting rotation angles (Gidaris et al., 2018; Zhai et al., 2019). Alternatively, co-training (Qiao et al., 2018) has been applied to semi-supervised image recognition where different models are trained on different "views" in order to learn complimentary information from the data. In general, most studies in the field focus on large-scale image recognition using 2D convolutional neural networks. Some works covered semantic segmentation (Hong et al., 2015; Papandreou et al., 2015), object detection (Misra et al., 2015; Tang et al., 2016) in 2D, as well as in graph-structured data analysis (Kipf and Welling, 2016).

Semi-Supervised Learning in Medical Imaging becomes a popular topic as large-scale datasets are made publicly available

[1] Authors contributed equally

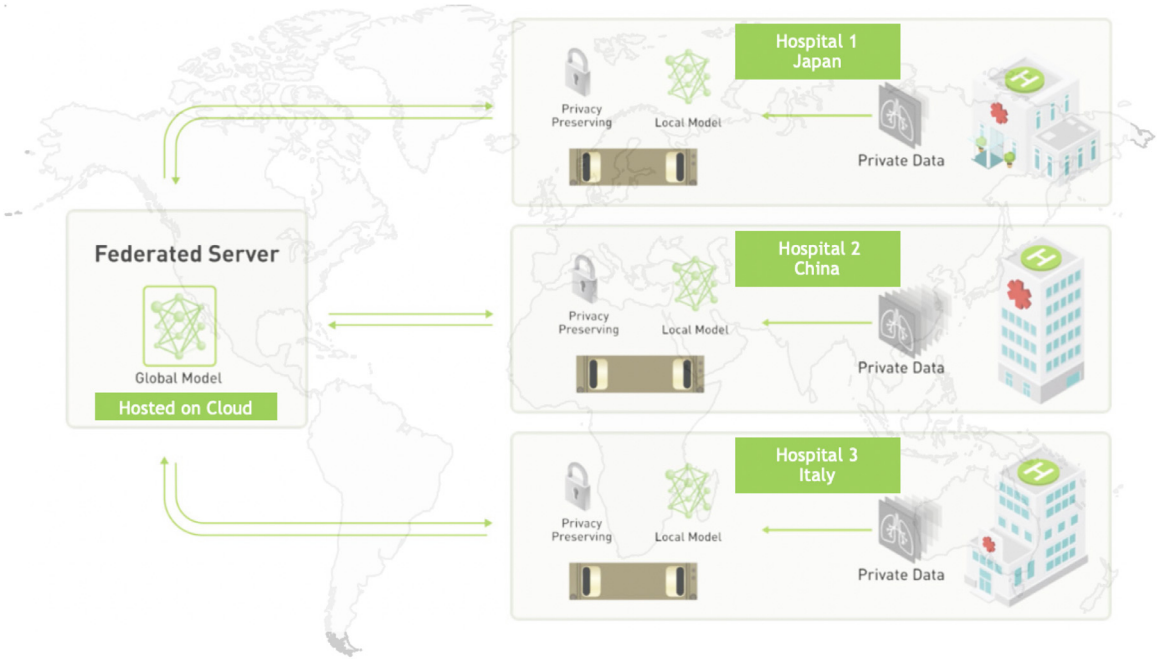

Fig. 1. Federated learning with privacy preserving in medical imaging. The central federated server communicates with clients from multi-national institutions by sharing weights and gradients of models, without exchanging any sensitive data information.

(e.g. Cheplygina et al. (2018)). But in the meantime, it is difficult to collect annotation for all datasets from experts or radiologists in practice. Bai et al. (2017) introduced a semi-supervised learning methods for cardiac image segmentation, using both deep neural networks and conditional random field (CRF). Li et al. (2018) proposed to add model regularization based on image rotation and flipping for unlabeled data in skin lesion segmentation.

Several works used self-ensembling and teacher-student interaction for medical image analysis, e.g. Liu et al. (2020a). Cui et al. (2019) proposed an adapted mean teacher model to improve accuracy of brain lesion segmentation leveraging both annotated and unannotated data. Yu et al. (2019) added an uncertainty-aware scheme to the teacher-student framework improving consistency regularization at training for left atrium segmentation in 3D MRI.

The multi-plane information from 3D medical images was used to enhance the supervised training model for prediction consistency in a co-training approach (Zhou et al., 2019). This method was further extended leveraging multi-view information of 3D medical images for full volumetric segmentation (Xia et al., 2020). Moreover, some researchers studied the usage of mixed supervision for medical image analysis (Shah et al., 2018; Mlynarski et al., 2019).

Federated Semi-Supervised Learning brings up challenges and complexity about how to exploit unlabeled data under a distributed learning setting (Jin et al., 2020). Unsupervised federated learning has been investigated for representation learning in a distributed setting (van Berlo et al., 2020). Federated self-learning was shown to be capable of detecting abnormality without any data label (Nguyen et al., 2019). Moreover, one-shot federated learning was introduced to conduct single-round communication between clients and server for both supervised and semi-supervised learning (Guha et al., 2019).

It is important to note that, in the context of federated learning, "semi-supervised learning" has a new dimension beyond the centralized training scenario, such as teacher-student model or co-training methods listed above. This new dimension is introduced by the multi-client setting and that different clients can have completely different annotation availability. We call it "global semi-supervision", as compared with "local semi-supervision" that is potentially viable for each individual client. As a matter of fact, the client-level network is compatible with other semi-supervised techniques, such as teacher-student model. In this paper, we propose an efficient and robust solution for the "global semi-supervision problem. To the best of our knowledge, there are very limited existing works on such federated semi-supervised learning or federated self-learning on unlabeled medical imaging data. In this paper, we propose a federated semi-supervised learning framework which utilizes unlabeled data to improve FL training and validate it in COVID region segmentation task. Our framework enables cross-institutional training on large-scale heterogeneous datasets without sharing sensitive private information.

3. Coronavirus affected region segmentation

Nowadays, machine learning based methods have been developed for medical imaging data acquisition, segmentation, and diagnosis of COVID-19 (Dong et al., 2020; Shi et al., 2020). Relying on the success of deep learning in medical image analysis, imaging characteristics of COVID-19 have been studied and analyzed from various perspectives. Some examples of affected regions of COVID-19 is given in Fig. 2. For the identification and segmentation of such regions, there are two major research directions using deep neural networks. The first one is to use classification models to distinguish normal subjects and patients. Class activation maps (CAM) can be extracted from these models that correspond to the affected region inside the lung area (Bai et al., 2020; Li et al., 2020b; Mei et al., 2020; Wang et al., 2020). The second direction is to apply 3D segmentation networks, typically fully convolution networks (FCN), and directly extract the COVID-19 affected regions following an image-to-image fashion (Fan et al., 2020; Huang et al., 2020; Liu et al., 2020b; Shan et al., 2020; Xie et al., 2020; Zhang et al., 2020; Zhou et al., 2020) shown in Fig. 3.

In this paper, we follow the second approach, using 3D U-shape FCN (Liu et al., 2018) as our baseline model, to segment the ground glass-like opaque (GGO) regions (COVID-19 affected re-

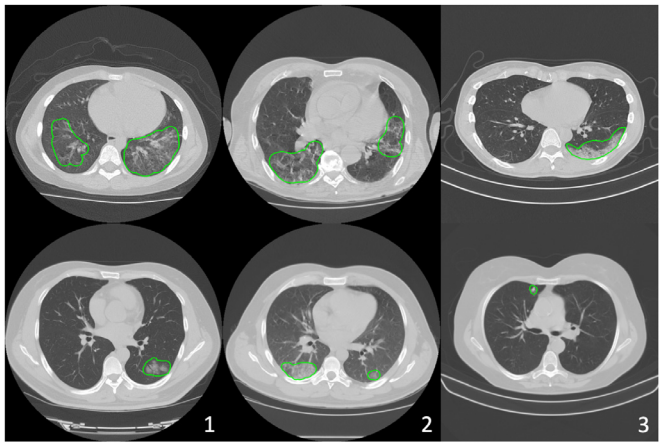

Fig. 2. The axial planes of chest CT scans from three different sites. Areas inside green contours represent COVID-19 affected regions annotated by radiologist. The appearance of the affected region identified as "infiltrates" range from diffused ground glass opacity (COVID, upper row) to focal nodules (lower row). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

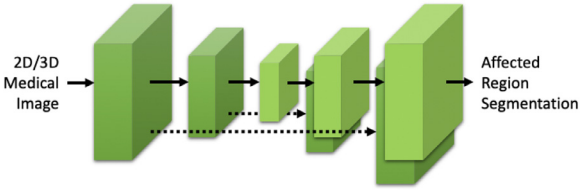

Fig. 3. The general architecture of fully convolutional network (FCN) for COVID-19 affected region in medical imaging. Dashed lines denote skip connections which feed earlier feature maps to later neural network layers.

gions, or COVID for short) of the lung from the 3D chest CT. However, it is noteworthy that our framework is independent of the exact neural network architecture used. This 3D U-shape FCN can be initialized with 2D model. Specifically, during the model initialization, weights of the 3D encoder are directly transferred from the pre-trained 2D ResNet with necessary operation conversion. For instance, 2D 3×3 convolutions are converted to 3D $3 \times 3 \times 1$ convolutions with the same amount of parameters. The parameters of batch normalization can be transferred directly without any modification. Each voxel is predicted as either foreground (COVID) or background. The output of models is a two-channel probability map after soft-max activation. The model is partially initialized with pre-trained weights of ResNet (He et al., 2016) from ImageNet classification for the encoder layers. We use the Adam optimizer to minimize the soft Dice loss (Milletari et al., 2016) given as

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_i y_i \hat{y}_i}{\sum_i (y_i)^2 + \sum_i (\hat{y}_i)^2}. \tag{1}$$

Here, y represents ground truth label and $\hat{y}$ is the prediction from the deep learning model.

Implementation Details. The CT volumes are converted to the most frequent resolution 0.8mm $\times$ 0.8mm $\times$ 5.0mm of the chest CT datasets before training and testing/inference. The actual input of the model is a cropped region-of-interest (ROI) with fixed size of $160 \times 160 \times 32$ during training. The patches sampled from the CT volumes are fed into the network for training. Patches are sampled with equal chance from foreground or background regions to maintain the training sample balance. The intensity of CT is clipped between Hounsfield units (HU) 0 and -1000, and mapped to the range of $[0, 1]$. The data augmentation strategy includes random flipping, random rotation, and random intensity shift. Moreover, the deep learning model is trained using 16GB NVIDIA Tesla V100 GPUs, with the pipeline developed using NVIDIA Clara Train SDK Platform (Clara, 2020). To achieve segmentation of the entire CT at inference, we use a crop size of $224 \times 224 \times 32$ to increase inference speed and to avoid artifacts caused by cropping. The sliding step is 16 along three axes. The batch size for training is 4, and learning rate is set to 0.0001.

4. Federated semi-Supervised learning

In this section, we introduce our framework, federated semi-supervised learning, for COVID region segmentation in 3D chest CT. The framework is designed to leverage unlabeled data for federated learning. In the following sub-sections, we explain the mechanism of our frameworks in details: the fundamentals of federated learning for COVID region segmentation are introduced in Section 4.1; then we illustrate the settings of federated semi-supervised learning in the segmentation task in Section 4.2; last but not least, the detailed implementation is further discussed in Section 4.3.

4.1. Federated learning

In our paper, the federated learning framework follows the conventional settings as (McMahan et al., 2016; Li et al., 2019). In the setting, a single server hosts the global optimal model at the moment, and meanwhile communicates with multiple clients. The neural network architecture is shared between server and clients. The communication is only about weight or gradient transferring, which is synchronized for all clients. All the data is associated with the clients. The data point from one client is invisible to both the server and other clients. There is usually no complicated computation on server side other than simple weighted aggregation. During FL training, the server collects "gradients" Δ_θ from clients simultaneously, aggregate those gradients, and send new model weights back to clients. The most important job for the server is **weight aggregation**. As shown in Eq. 2, the aggregation is conducted as weighted summation.

$$\hat{\Delta}_\theta \leftarrow \hat{w}_i \cdot \sum_{i=1}^{\mathcal{C}} \Delta_\theta^i \tag{2}$$

Here, $\mathcal{C}$ is the number of clients. We firstly weigh clients by n, the quantity of iterations per synchronization round. Because the model trained with more steps tends to have larger difference with the initial model, the updates from clients should be normalized to have the same update pace, which would avoid the unnecessary bias towards any one of the clients. Therefore, besides weights from iterations, we have additional weights $\mathcal{W}$ for each client assigned by users. Hence, the overall weight of each client contains two components as follows.

$$\hat{w}_i \leftarrow \frac{n_i}{\sum_i n_i} \cdot w_i, w_i \in \mathcal{W}, i = 1, \cdots, \mathcal{C} \tag{3}$$

Here, n_i is the iteration number per round of client i. Once aggregation is accomplished, the server sends out the updated model weights back to each client at the same time.

Next, the clients collect models weights from the server, fine-tune the model with their local data, and send out the new gradient to server. They are independent instances, which are not direction connection between each other. In this paper, the clients launch training jobs described in Section 3 for COVID region segmentation. Each client has its own chest CT data, and GPUs as computing resource. Moreover, the optimal model checkpoint for each client is selected based on its own validation set.

The server is launched first and the clients are initialized accordingly with the global model from server. After several FL rounds of server-client communication, the global model on the server would be improved with greater generalizibility, and the

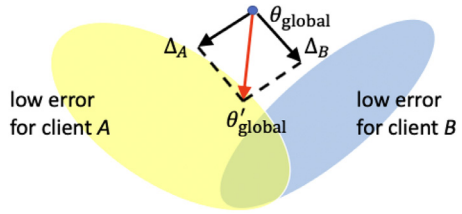

Fig. 4. FL is to find global low-error space of θ_{global} for client A and B after aggregation of "gradients" Δ_A and Δ_B. The low-error space is defined by each client, which could correspond to high accuracy, or high model consistency for self-supervision.

performance of local models would be further boosted. This intuition is shown in Fig. 4. The entire FL algorithm is shown in Algorithm 1 and is potentially feasible for all deep learning related applications of medical image analysis.

Algorithm 1 Federated learning for COVID region segmentation using weighted federated averaging.

Input:~number of clients $\mathcal{C}$, amount of global synchronization rounds $\mathcal{T}$, aggregation weights $\mathcal{W}$ for all clients, learning rate λ_i for each client i (for simplicity, we show the gradient descent update rule; not Adam optimization).
Output:~optimal θ_c for each client c.

procedure ServerUpdate
  initialize global model θ_0
  send θ_0 to all $\mathcal{C}$ clients
  for round $t = 1, 2, \cdots, \mathcal{T}$ **do**
    wait for all $\mathcal{C}$ to finish update
    $\left(\Delta^i_\theta, n_i\right) \leftarrow \texttt{ClientUpdate}(\theta_{t-1})$, $i = 1, \cdots, \mathcal{C}$
    $\hat{w}_i \leftarrow (n_i / \sum_i n_i) \cdot w_i$, $w_i \in \mathcal{W}, i = 1, \cdots, \mathcal{C}$
    $\hat{\Delta}_{\theta,t-1} \leftarrow \sum_{i=1}^{\mathcal{C}} \hat{w}_i \cdot \Delta^i_\theta$ ▷ aggregate updates
    $\theta_t \leftarrow \theta_{t-1} + \hat{\Delta}_{\theta,t-1}$
  end for
end procedure

function ClientUpdate(θ)
  $n \leftarrow$ amount of total training iterations
  $\theta_{\text{init}} \leftarrow \theta$
  for iteration $j \in 1, 2, \cdots, n$ **do**
    $\theta \leftarrow \theta - \lambda_i \nabla \mathcal{L}(\theta, j)$ ▷ optimize loss function
  end for
  $\Delta_\theta \leftarrow \theta - \theta_{\text{init}}$
  return (Δ_θ, n)
end function

4.2. Federated semi-Supervised learning on COVID region segmentation

We assume that some of clients may not have enough expertise to create accurate annotation for their datasets. But the information of their patient data is still valuable to the rest of community. In practice, some clients only possess unlabeled chest CT images, which could be from COVID-19 patients, pneumonia patients, or normal subjects. However, each of those database has its own merits for jointly training COVID-19 affected region segmentation. For instance, COVID-19 related CT images contain informative context from its appearance inside the lung region. Non COVID-19 related CT images can provide guidance on false positive removal for the COVID-19 affected region segmentation model (ideally nothing should be predicted based on those CT images). In order to utilize the clients' side unlabeled data, we propose a novel framework of federated semi-supervised learning for COVID-19 affected region segmentation.

Under the existing FL setting, we create a unsupervised client, which only has unlabeled data, without modifying the FL server. The unsupervised clients retrieve the global model $\mathcal{H}(\cdot;\theta)$ with weights θ from the server and apply it to their own database. Then, we would like to enforce consistency of predictions from the global model, to further adjust model weights. Similar to (Berthelot et al., 2019b; 2019a; Sohn et al., 2020), we introduce a new loss function based on data augmentation. The assumption behind this is that the generalizable model should perform the same with original data and slightly perturbed data. We assume that the perturbation can still be within the range of the actual data distribution.

Let u denote one CT image from the unlabeled database $\mathcal{U}$. We create the pseudo-label $\bar{y}$ using prediction of the current model with input u and hard thresholding (against 0.5).

$$\bar{y} = \begin{cases} 1, & \mathcal{H}(u) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

In order to train the global model locally, the new loss function minimizes the difference between pseudo-label $\bar{y}$ and the prediction after augmentation $g(\cdot)$:

$$\mathcal{L}_{\text{consistency}} = \mathcal{L}(\bar{y}, \mathcal{H}(g(u))). \tag{5}$$

Here, $\mathcal{L}$ is the soft Dice loss for segmentation tasks, as in Eq. 1. Other loss functions, such as cross entropy or ℓ_2 loss, can also be used here. Data augmentation could be random scale shift, Cutout (DeVries and Taylor, 2017), adding Gaussian noise, and so on. In general, $g(\cdot)$ on image appearance should not generate out-of-distribution samples. Because the pseudo-label is generated fully from the image, the ambiguous area (output probability close to 0.5) in the prediction may not be helpful for model training. To further improve the quality of pseudo-label, a confidence threshold $\tau \in [0.5, 1]$ can be added to determined the regions used for computing the loss:

$$\mathcal{L}_{\text{consistency}} = \mathbb{1}(\mathcal{H}(u) > \tau)\mathcal{L}\left(\hat{y}, \mathcal{H}(g(u))\right). \tag{6}$$

Comparison with Centralized Semi-Supervised Learning. After training several epochs, the server collects gradients Δ_θ from both supervised clients $\mathcal{C}_s$ and self-supervised/unsupervised clients $\mathcal{C}_u$ for model aggregation:

$$\hat{\Delta}_\theta \leftarrow \sum_{i \in \mathcal{C}_s} \hat{w}_i \cdot \Delta^i_\theta + \sum_{i \in \mathcal{C}_u} \hat{w}_i \cdot \Delta^i_\theta. \tag{7}$$

Although it seems to be similar with jointly loss training in centralized semi-supervised learning (SSL), the proposed FL approach has intrinsic differences compared to SSL. First, the objective function $\mathcal{L}$ is only visible to each client itself. The gradients of clients may diverge to different local minima, which is also known as weight divergence problem in FL. Second, for the communication efficiency of FL training, the model update cannot be conducted per local iteration. It makes the federated semi-supervised learning even more challenging. The weights $\hat{w}_i$ between different clients, and the learning rate λ_i of each clients may play an important role for the final model performance. Third, due to the difference of patients population, scanning protocol, and scanners, the data distribution of clients are usually non-identical and non-independent (non-iid) in the application. For instance, one client has data from patients at early stage, and another client may possess data with severe conditions only. There is clear appearance/domain difference between these two clients. It is unclear how to handle such domain difference under federated semi-supervised learning setting. All these factors make federated semi-supervised learning much more difficult comparing to centralized semi-supervised learning.

The federated semi-supervised algorithm for clients with unlabeled data is shown in Algorithm 2, which is also feasible for

Algorithm 2 Self-supervised learning algorithm at clients with unlabeled data for COVID region segmentation.

Input:~learning rate λ for each client, unlabeled data pool $\mathcal{U}$, global model $\mathcal{H}$.
Output:~weight difference Δ_θ.

function SELFSUPERVISEDCLIENTUPDATE(θ)
  $n \leftarrow$ amount of total training iterations
  $\theta_{\text{init}} \leftarrow \theta$
  for iteration $j \in 1, 2, \cdots, n$ **do**
    $\hat{u}_j \leftarrow \text{Augment}(u_j)$, $u_j \in \mathcal{U}$ ▷ input perturbation
    $\mathcal{L}(\theta, j) \leftarrow \mathcal{L}_{\text{consistency}}\big(\mathcal{H}(u_j; \theta), \mathcal{H}(\hat{u}_j; \theta)\big)$
    $\theta \leftarrow \theta - \lambda \nabla \mathcal{L}(\theta, j)$ ▷ optimize loss function
  end for
  $\Delta_\theta \leftarrow \theta - \theta_{\text{init}}$
  return (Δ_θ, n)
end function

all machine learning related applications of medical image analysis. Since the segmentation model requires necessary supervision to train, at least one client needs to possess labeled data in training. However, federated unsupervised learning would be a future direction to explore for representation learning.

4.3. Implementation details

The FL implementation for COVID-19 affected region segmentation is constructed with NVIDIA Clara Train SDK using TensorFlow 1.14 (Clara, 2020), which uses the gPRC protocol for communication between the server and clients during model training. Since we use patch-based training strategy for 3D images (because of GPU memory limited and efficient training), we modify the client training with the same iterations per epoch, and the same epochs per round. Thus, the contributions from different clients are equivalent if the clients' weights are the same.

Training-from-scratch for unsupervised clients are not meaningful because no guidance from the global model or local unlabeled data is available. This issue may be mitigated after several rounds of model aggregation, but it slows down the overall FL training efficiency. Therefore, to enable reasonable training of the unsupervised client, for the federated learning with 1 supervised and 1 unsupervised clients, we first train the supervised client for 500 epochs (on the data of the supervised client). The federated learning then starts with this model and gets trained for another 500 epochs, such that the unsupervised client can start with a meaningful feature representation. For cases with two supervised clients, the federated learning starts from scratch and gets trained for 1000 epochs.

Each epoch contains 20 iterations, and each round of synchronization contains 20 epochs. To be specific, we set the same iteration numbers per federated round for all clients to ensure the similar training paces of clients and mitigate potential side-effects caused by client asynchronization. Since the clients with more training iterations (or optimization steps) tend to converge faster than the one with less iterations. Validation is conducted also every 20 epochs for supervised clients. Then the optimal model checkpoint is determined using the validation dataset on supervised clients.

Moreover, the learning rate is $5e^{-6}$ for unsupervised clients. It cannot be as large as the one in supervised clients because large learning rate could cause the model to overfit the self-supervised tasks quickly, potentially diverging the gradient towards a biased direction. The same assumption can be made for client weights. The weights of unsupervised clients cannot be larger than ones in supervised client, otherwise the convergence of global segmentation model becomes slow and unstable. The actual input of the model is a cropped region-of-interest (ROI) with fixed size $160 \times 160 \times 32$ for training. The patches sampled from the CT volumes are fed into the network for training, and they are sampled randomly over the entire CT volume. And the intensity of CT is clipped between Hounsfield units (HU) 0 and -1000, and mapped to the range $[0, 1]$. Adam optimizer is used to minimize consistency loss.

5. Experiment and results

5.1. Data and expert annotation

COVID-19 Population: patients undergoing CT evaluation with SARS-CoV-2 infection confirmed by RT-PCR were identified from three international centers: 1) 736 scans of 700 patients from the First Affiliated Hospital of Hubei University of Medicine in Hubei Province, China (referred to as **Image_1**), 2) 496 scans of 244 patients from the Self-Defense Forces Central Hospital, Tokyo, Japan (referred to as **Image_2**), and 3) 472 scans of 147 patients from San Paolo Hospital, Milan, Italy (referred to as **Image_3**). It is important to note that the image acquisition from these three institutions varies considerably: in China, the CT scans were routinely obtained on the same day as a positive RT-PCR in an acute setting during the initial outbreak period; in Japan, the patients were a mixture of incidental Diamond Princess cruise ship exposures or community acquired COVID-19, with a diverse multinational population; and in Italy, CT scans varied from acute care screening to inpatients, commonly later in the disease process. Therefore, the conditions included in this study are fairly diverse, and we observed domain shifts among the three cohorts, as shown in Fig. 2.

Control Population a control population was identified from one institution and one publicly available dataset: 1) 38 scans of 38 patients at the National Institutes of Health undergoing CT evaluation of known non-COVID-19 pneumonias from bacteria, fungi, and non-COVID viruses were included as a "other pneumonia" cohort (referred to as **Image_P**), 2) 101 images of 101 patients with unremarkable lung findings from men with prostate cancer at the National Institutes of Health were included as a non-diseased "normal" cohort (referred to as **Image_N**), and 3) a total of 474 scans of 474 patients were derived from the publicly available LIDC dataset (Armato III et al., 2011) consisting of lung nodule data (referred to as **Image_LIDC**).

Annotation CT scans underwent a centralized evaluation by two expert radiologists for confirmation and localization of lung disease patterns related to COVID-19. Regions of CT infiltrates were manually delineated using ITK-SNAP tool (Yushkevich et al., 2006). The most common "CT infiltrate" was ground glass opacities, followed by consolidation. To simulate the uneven distribution of annotation resource, out of the entire dataset, 671 (out of 736), 88 (out of 496), and 186 (out of 472) scans were annotated for Image_1, Image_2, and Image_3, respectively. Note that here, the three datasets have different degree of inter-observer variance from experts: all data of Image_1 and Image_2 are annotated by the same expert radiologists in the same institute, while Image_3 is annotated by different radiologists from 2 different countries.

5.2. Analysis

Data Split and Experiment Design Among the images with experts' annotation, we split the dataset randomly into training/validation/testing sets. To ensure sufficient testing and validation data, the following splits are used: 447/112/112 for Image_1, 30/29/29 for Image_2, and 124/31/31 for Image_3. To test the system's performance under an unbalanced situation, for most experiments, we use two of the three datasets as two clients for feder-

Table 1
Accuracy (Dice's score) comparison with different experimental settings. The upper part of the table is for regular training with 1 and 2 GPUs, the bottom part is for federated learning with two clients. "**Sup.**" indicates the supervised client, and "**Unsup.**" indicates the client which contributes unlabeled data only.

		Image_1		Image_2		Image_3	
		Valid. Acc.	Test Acc.	Valid. Acc.	Test Acc.	Valid. Acc.	Test Acc.
Image_1 (1 GPU)		0.571±0.005	0.575±0.003	-	0.578±0.013	-	0.577±0.011
Image_2 (1 GPU)		-	0.479±0.008	0.626±0.004	0.625±0.005	-	0.536±0.010
Image_3 (1 GPU)		-	0.480±0.010	-	0.570±0.015	0.649±0.004	0.607±0.007
Image_1, Image_2 (1 GPU)		0.572±0.003	0.578±0.006	0.634±0.005	0.593±0.011	-	0.579±0.013
Image_1, Image_2 (2 GPUs)		0.581±0.004	0.601±0.004	0.639±0.007	0.594±0.010	-	0.575±0.009
Image_1, Image_3 (1 GPU)		0.564±0.003	0.574±0.003	-	0.580±0.017	0.636±0.005	0.574±0.015
Image_1, Image_3 (2 GPUs)		0.575±0.008	0.586±0.008	-	0.605±0.001	0.651±0.007	0.600±0.008
Image_2, Image_3 (1 GPU)		-	0.489±0.008	0.627±0.006	0.600±0.007	0.650±0.008	0.615±0.010
Image_2, Image_3 (2 GPUs)		-	0.513±0.011	0.640±0.003	0.613±0.007	0.665±0.004	0.638±0.005
Image_1, Image_2, Image_3 (1 GPU)		0.565±0.010	0.572±0.010	0.623±0.006	0.579±0.014	0.637±0.008	0.575±0.016
Image_1, Image_2, Image_3 (2 GPUs)		0.577±0.006	0.590±0.008	0.642±0.002	0.608±0.008	0.647±0.006	0.591±0.008
FL Client 0	FL Client 1						
Sup., Image_1	Sup., Image_2	0.566±0.002	0.573±0.005	0.637±0.016	0.603±0.019	-	0.579±0.010
Sup., Image_1	Unsup., Image_2	0.563±0.004	0.569±0.006	-	0.590±0.011	-	0.568±0.011
Unsup., Image_1	Sup., Image_2	-	0.478±0.007	0.628±0.006	0.622±0.004	-	0.553±0.015
Sup., Image_1	Sup., Image_3	0.549±0.010	0.552±0.009	-	0.566±0.023	0.628±0.018	0.560±0.028
Sup., Image_1	Unsup., Image_3	0.561±0.003	0.564±0.001	-	0.569±0.010	-	0.551±0.010
Unsup., Image_1	Sup., Image_3	-	0.458±0.010	-	0.527±0.032	0.637±0.008	0.579±0.023
Sup., Image_2	Sup., Image_3	-	0.489±0.007	0.616±0.011	0.588±0.017	0.664±0.004	0.605±0.010
Sup., Image_2	Unsup., Image_3	-	0.471±0.014	0.627±0.004	0.625±0.008	-	0.541±0.017
Unsup., Image_2	Sup., Image_3	-	0.461±0.004	-	0.551±0.019	0.637±0.009	0.586±0.019

ated learning, while the other is set up as an "unseen" domain in order to test the models' generalizability.

Supervised Baselines The baselines for our study is the model learnt under fully supervised conditions, this includes models trained with single source datasets, as well as centralized training with mixed Image_1, Image_2, and Image_3 datasets. Here, three combinations of two sites are performed matching the experiments of federated learning, where we always want to keep an unseen domain so as to evaluate the generalizability of the trained model.

Further note that for fair comparison with federated learning where two sites train on their own GPU, besides training with 1 GPU, we also added a study with 2 GPU training. Also to reduce the influence from training randomness, each baseline is repeated five times under deterministic training with different random seeds, and the mean performance with standard deviation is reported.

Parameter Setting The setting of clients was chosen empirically with some trials of different combinations and selecting the ones with the best performance as following: hard Dice loss on foreground with $\tau = 0.90$, supervised client learning rate 1e-4, unsupervised client learning rate 5e-6. The augmentation for computing consistency loss is random scale shift of intensity.

As shown in Table 1, single-site training suffers from generalizability on other domains, models trained on Image_2 and Image_3 has 10% Dice drop for the testing accuracy on Image_1 comparing to the model trained on Image_1, and similar patterns can be observed from the other two scenarios.

The performance of a model on a set of data depends on: 1) the capability and generalizability of the model, 2) the data distribution and similarity between the training and testing data, and 3) the annotation variation for evaluation. Therefore, "unseen domain" does not necessarily mean "domain with significant shift", and furthermore, "domain with significant shift" does not necessarily mean "domain where the model will see significant performance drop". Meanwhile, the "performance" also has two dimensions: 1) absolute performance with the same model on different datasets; and 2) relative performance on the same data with different models. Absolute performance is hard to predict, while in most cases, relative performance with model trained on the same data has the highest score. In other words, the following holds for most cases: Two datasets 1 and 2, let's denote the performance of the model trained on i and tested on j as p_{ij}, then we have: for absolute performance p_{11} may or may not $> p_{12}$, similarly, p_{22} may or may not $> p_{21}$; on the other hand in most cases for relative performance, $p_{11} > p_{21}$, and $p_{22} > p_{12}$. This can be observed from Table 1: first row, absolute performance $p_{11} < p_{12}$; while fourth column, relative performance $p_{22} > p_{12}$.

Since the performance and robustness of a model is largely determined by the training data, if the supervised client is "strong", i.e. containing sufficient amount of data (in our case Image_1), the trend of adding more data (Image_2/3), whether supervised or unsupervised, is fairly stable, and we list the observations below.

Combining datasets in a centralized training can help promoting both the accuracy and the generalizability of a model. The gap of data distributions from two datasets is mitigated via randomized mini-batching. Thus, models trained with data centralization perform better than the models on Image_1 or Image_2 independently. Since FL is also capable to collect information from multiple datasets, FL trained models with both supervised clients have comparable performance to single-GPU trained model on Image_1 or Image_2.

Federated semi-supervised learning has approximately 1% Dice's score drop compared to supervised FL. However, the model performs better than the model trained on Image_1, Image_2, or Image_3 independently in general (see the testing results on Image_2 and Image_3). It demonstrates that **the unlabeled data from different clients are valuable to train a generalizable model**. The visual results are shown in Fig. 5, and our model predicts segmentation masks with better shape and less false positives.

On the other hand, if the supervised client is not "strong" or less representative of overall data distribution, i.e. containing limited amount of data (in our case Image_2), or containing annotation variance/bias (in our case "Image_3"), then the trend of adding more data can become unstable.

First, federated learning with "Image_1" and "Image_2" meets our expectation using both fully- or semi-supervised setting. The supervised clients' models maintains the same level performance, and the model generalizability has been further improved over three datasets. Second, the dataset "Image_3" seems to possess bias regarding to overall data distribution. Adding "Image_3" into training may down-grade the performance a bit (line "Unsup., Im-

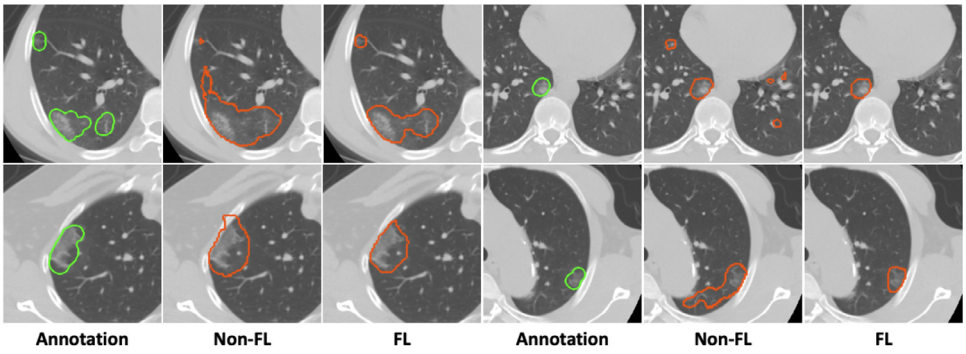

Fig. 5. Visualizations of federated semi-supervised segmentation of COVID regions in 3D CT (from the testing set of unsupervised client **Image_2**). "Non-FL" indicates results from the model trained with Image_1 along, and "FL" denotes results from the model trained with federated semi-supervised learning on Image_1 and Image_2. The segmentation results using the proposed framework captures the ground truth shapes better and has less false positives.

age_2 Sup., Image_3" in Table 1), which is potentially caused by biased annotation protocols from different radiologists of different countries. However, unsupervised "Image_3" client (appearance of "Image_3") is still able to help supervised client "Image_2" to improve its model generalizability (line "Sup., Image_2 Unsup., Image_3" in Table 1). Moreover, due to much larger size of "Image_1" compared to "Image_2" and "Image_3", unsupervised client "Image_1" would down-grade its own performance by a large margin compared to its supervised counter-part.

Because federated learning is one type of distributed learning using multiple computing units. The centralized 2-GPU training was conducted to make further comparison. From Table 1, we can see the 2-GPU centralized training performs much better than other setting. It is mainly caused by multi-GPU training with larger batch size per iteration. The supervision from both dataset is stronger and pulls the model weights towards a stable condition.

5.3. Ablation studies

Several components are configurable under the proposed federated semi-supervised learning framework, specifically, the image generation and loss functions for the unsupervised clients, the learning rate of each client, and the aggregation frequency for the server. The default FL framework followed the setting described above. We used same random seeds for all ablation study experiments. Due to the high amount of ablation studies, we trained a single model for each configuration instead of repeating 5 times.

Loss functions of self-/un-supervised learning In addition to foreground hard Dice loss, We experimented with another nine potential loss functions for $\mathcal{L}_{\text{consistency}}$, including L1, L2, cross entropy (CE), and hard and soft Dice (Dice_H and Dice_S), all losses have two configurations: whole image and foreground (_F). Table 2 summarizes their performance. As shown in the table, the hard Dice generally has an edge over other losses, while using foreground only may not yield better performance.

Data augmentation We further conduct comparison against different augmentation strategies for unsupervised clients. "Gauss_x" means adding voxel-wise Gaussian noise into CT volumes with zero mean and variance x. From Table 2, higher noise level increases generalizability of the model globally. "Cutout_1" DeVries and Taylor (2017) means masking a random cube (size $20 \times 20 \times 3$) from the re-sampled image with zero, and "Cutout_2" means masking five of such random cubes from the re-sampled image with zero. The performance drops significantly when increasing the "Cutout" regions in augmentation (shown in results on Image_3). "Fix_y" follows the idea of "FixMatch" (Sohn et al., 2020) to create both strongly and weakly augmented samples for unsupervised learning. And the pseudo-label is generated based on the weakly augmented samples. y means the level of random intensity shift for weakly augmented samples, and $1.0 - y$ is for strongly augmented samples. $y = 0.1$ creates larger gap between strongly and weakly augmented samples, compared to $y = 0.25$. Clearly from the Table 2, the larger difference of samples corresponds to the more generalizable models. As for a particular task and data, experiments will be needed to select the best data augmentation strategies for computing consistency loss.

Aggregation weights During each round of server aggregation, updates from different clients can be weighted before they are aggregated together to update the model on the server end. Here, we adjust the weight of the unsupervised client from 0.1 to 1.0. As shown in Table 2, the behavior of this parameter is not linearly correlated with the performance on specific datasets. This may be

Table 2
Ablation studies for unsupervised loss, augmentation functions of unsupervised client, and aggregation weights.

	Image_1		Image_2	Image_3
	Sup.		Unsup.	Unseen
	$\text{Acc}_{\text{Valid.}}$	Acc_{Test}	Acc_{Test}	Acc_{Test}
Baseline	0.562	0.571	0.546	0.556
Loss Function				
L1	0.540	0.552	0.585	0.571
L1_F	0.553	0.548	0.552	0.547
L2	0.549	0.552	0.579	0.548
L2_F	0.540	0.539	0.575	0.548
CE	0.557	0.560	0.593	0.564
CE_F	0.559	0.571	0.535	0.545
Dice_S	0.565	0.567	0.592	0.565
Dice_SF	0.552	0.561	0.576	0.546
Dice_H	**0.574**	0.569	**0.596**	0.569
Data Augmentation				
Gauss_0.1	0.550	0.552	0.557	0.566
Gauss_0.9	0.548	0.551	0.578	0.560
Cutout_1	0.548	0.548	0.569	0.543
Cutout_2	0.547	0.556	0.555	0.529
Fix_0.1	0.554	0.559	0.591	0.556
Fix_0.25	0.552	0.549	0.572	0.547
Aggregation Weights				
1.0:1.0	0.562	0.571	0.546	0.556
1.0:0.75	0.560	**0.573**	0.574	0.553
1.0:0.5	0.563	0.562	0.588	0.570
1.0:0.25	0.562	0.558	0.592	0.570
1.0:0.1	0.563	0.568	0.572	0.554
Partial Weight Update				
Final	0.535	0.550	0.598	0.555
Block_10	0.557	0.564	0.573	0.564
Block_6	0.564	0.572	0.592	**0.578**

Table 3
Ablation studies for unsupervised client datasets and another base network.

	Image_1		Image_2	Image_3
	Sup.		Unseen	Unseen
	$Acc_{Valid.}$	Acc_{Test}	Acc_{Test}	Acc_{Test}
N	0.548	0.546	0.588	0.550
LIDC	0.550	0.564	0.593	0.562
P	0.566	0.562	0.591	0.556
SegResNet	0.533±	0.540±	0.596±	0.489±
Image_1	0.002	0.004	0.003	0.006
SegResNet	0.534	0.543	0.607	0.512

explained by the fact that the updates learnt from unsupervised client, while catching characteristics of Image_2, also contains fair amount of noise due to its unsupervised nature. Therefore, reducing the weight from unsupervised client leads to not only the reduction of the influence from Image_2, but also the noise from it.

Partial model update Another interesting experiment is to share partial weights only across clients. In Table 2, "final" denotes the last convolutional layers are not shared between clients, "block_10" denotes the last DenseBlock and final convolutoinal layers together are not shared, and "block_6" illustrates all layers after encoder (equivalent to decoder path plus last convolutional layer, blocks 6, 7, 8, 9, 10 + last convolutional layer) are not shared. From the comparison, the less layers are shared, the better performance the global model has. The FL model, with encoder-only sharing, possess excellent performance with better generalizability for both seen and unseen datasets. And such finding might be caused by intrinsic difference between supervised and unsupervised tasks. Federated encoder training could jointly learn a good feature representation across multiple clients' database, then the segmentation task is handled better with independent decoder with labeled data. Meanwhile, the similar strategy could be applied even for FL with all supervised clients, or multi-task FL. Furthermore, the partial weight sharing provides safer solutions for privacy preserving, since only partial model information is used in client-server communication. The experimental results in the paper, other than ones in this ablation study, are federated averaging with the full models (aggregating all model weights).

Other datasets and other network The control population of Image_N, Image_LIDC, and Image_P can also serve for the unsupervised client. From the cohort relationship perspective, Image_N is farthest from the candidate COVID-19 data, since it contains almost no abnormal regions. Image_LIDC may have certain similarity, since some COVID-19 cases can have nodule-like focal consolidation areas. Image_P would be the most close to COVID-19 dataset, since COVID is common for scans of pneumonia. As illustrated from the results listed in Table 3, the above relationship can be observed from the validation accuracy, which is the most predicable. The testing accuracy for Image_1 also shows a similar pattern, though Image_LIDC and Image_P have similar performance. The testing accuracy on the two unseen domains are more unpredictable. Besides the current base network (Liu et al., 2018), other networks serving similar purpose can also be used. We did an experiment with the SegResNet proposed in (Myronenko, 2019). The single site training on Image_1 is also trained five times, and the federated semi-supervised learning is setup with the same configurations of loss, etc. as previous baseline study. As shown in Table 3, the overall accuracy is not as high as the current network, but the federated result is slightly better than Image_1 alone. Again as mentioned previously, our framework is flexible to host most networks as the base network.

Learning rates We studied different learning rates (LR) on the unsupervised client, while LR of the supervised client is fixed at

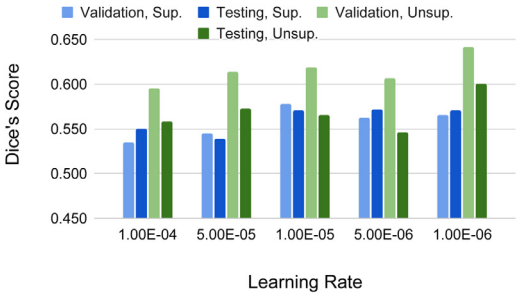

Fig. 6. Accuracy (Dice's score) comparison of different learning rates of un-/unsupervised clients.

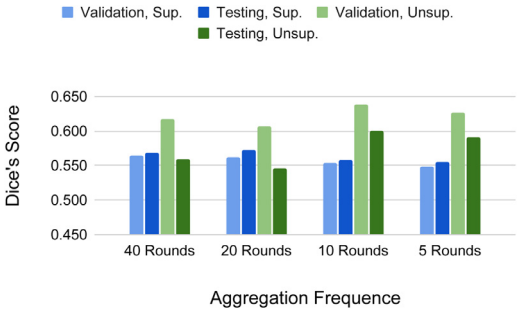

Fig. 7. Accuracy (Dice's score) comparison of different aggregation frequency (per 5, 10, 20, 40 rounds).

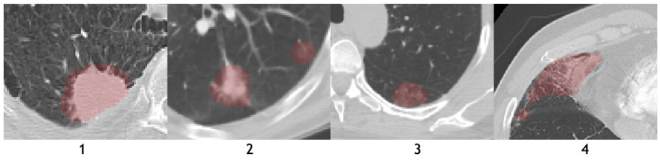

Fig. 8. Model performance on LIDC dataset, capturing solid nodule (1, 2); ground glass nodule (2, 3); and other abnormal patterns (4).

1e-4 for fair comparison. The value of LR varies from 1e-6 to 1e-4. And LR of the unsupervised clients cannot be too large or even larger than the one used in the supervised clients. In the FL setting, each round of training is client-independent. Therefore, the model weights would quickly converge to over-fit the self-supervised tasks when LR is large. It indicates that the over-fitted model weights might be biased towards unexpected directions, which would make the overall training procedure unstable. Such side-effect is shown in the Fig. 6. Here, the supervised client is using Image_1, and unsupervised client is using Image_2. The accuracy on the validation/testing set of unsupervised client is lower, when LR is set to a higher values. Higher LR also affects the performance on the supervised client, which demonstrates that 1) FL training become less effective; 2) the correlation between self-supervised tasks and supervised tasks is weak. Thus, lower LR on unsupervised client generally benefits all clients in FL. **Aggregation frequency** Another question for FL is how often clients and server should communicate to each other. In other words, how often should client model weights be aggregated. Ideally, if the model aggregation happens every training iteration, then federated semi-supervised learning is equivalent with standard semi-supervised learning with equal sampling chance from both labeled and unlabeled datasets. In order to verify the effects of aggregation frequency, we conducted the ablation study with aggregation per 5, 10, 20, 40 epochs shown in Fig. 7. Here, supervised client is using Image_1, and unsupervised client is using Image_2. In general, when the aggregation is more frequent, the performance on the self-supervised side improves (see the performance on the test-

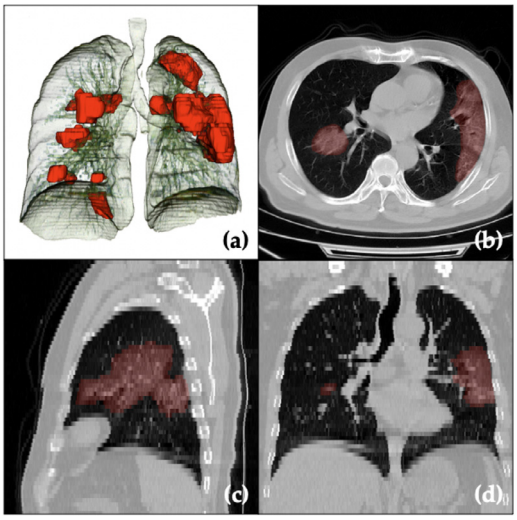

Fig. 9. Visualization of the COVID-19 affected region segmentation/prediction (red regions) together with lungs and airways (a) in 3D space, and (b,c,d) in different (axial, sagittal, coronal) planes of a raw CT image. Note that slice thickness is 5 mm (as compared with 0.8 mm in-plane), which is the case for most images in this work. . (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ing data of the unsupervised client). At the same time, the performance of the supervised clients become slightly worse. More frequent aggregations correct the training trajectory on unsupervised clients more often, and the overall training becomes smoother. In reality, it is not practical to synchronize clients frequently due to the bandwidth limitations. The trade-off between the aggregation frequency and the communication cost needs to be tuned for the optimal training efficiency depending on the specific applications.

6. Conclusions and discussion

In this paper, we proposed a *federated semi-supervised learning* framework for COVID-19 affected region segmentation in 3D chest CT (3D visualization with airway and lung in Fig 9). The proposed framework is capable to grasp valuable information from the clients which only have unlabeled data. Meanwhile, the privacy of all patients has been preserved, and they do not need to share their own database for collaborating on joint model training. Moreover, after jointly training with supervised and un-/unsupervised clients, the generalizability has been improved for not only each client's database, but also on the unseen data domain. We found out even the client with pure non-COVID database is able to help model training for COVID-19 affected region segmentation via false alarm rejection.

One thing to further clarify is that the aim of work is to segment the disease affected regions that is reflected in CT images, and the annotation is solely based on dataset consisting of COVID-19 cases. The "COVID-19 affected region" is identified as the abnormal regions in the context of these cases. Therefore, the model is not trained to discriminate against other type of abnormalities, e.g. other pneumonia or cancer. From a "pipeline" point of view, additional classification to tell the difference can follow the proposed segmentation method, but will need additional data and annotations. To give an example, we used our trained network to perform inference on LIDC dataset Armato III et al. (2011). Fig. 8 showed four different cases of abnormalities captured by our model on LIDC images: 1. solid nodule, 2. mixed solid and ground glass nodule, 3. ground glass nodule, and 4. other abnormal pattern. As nodules have similar appearance in CT as abnormalities caused by COVID-19, those regions are detected.

Our proposed framework has been validated on a multi-national database cohort with population, equipment, and demographic variation. In addition, comprehensive ablation studies have been explored for the proposed framework.

The proposed *federated semi-supervised learning* framework is general for machine learning based applications of medical image analysis. Given the limited literature, this work may initiate a promising direction for the future study of medical image analysis. Still, there are open questions along this research direction. For example, there are potential domain gaps between supervised and unsupervised clients. It is an unsolved problem of how to better model this domain gap and mitigate it during federated learning. Another example is how to adaptively aggregate contributions from different clients based on the quality and and not just the quantity of a clients' database. Because there are a lot variables in the semi-supervised framework, and complexity is even higher compared to regular FL, we hope our work is a good starting point for future exploration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Dong Yang: Conceptualization, Methodology, Software, Visualization, Validation, Formal analysis, Writing - original draft. **Ziyue Xu:** Conceptualization, Methodology, Software, Visualization, Validation, Formal analysis, Writing - original draft. **Wenqi Li:** Conceptualization, Writing - review & editing. **Andriy Myronenko:** Conceptualization, Writing - review & editing. **Holger R. Roth:** Conceptualization, Writing - review & editing. **Stephanie Harmon:** Methodology, Writing - review & editing. **Sheng Xu:** Methodology, Writing - review & editing. **Baris Turkbey:** Methodology, Writing - review & editing. **Evrim Turkbey:** Methodology, Writing - review & editing. **Xiaosong Wang:** Methodology, Writing - review & editing. **Wentao Zhu:** Methodology, Writing - review & editing. **Gianpaolo Carrafiello:** Methodology, Writing - review & editing. **Francesca Patella:** Methodology, Writing - review & editing. **Maurizio Cariati:** Methodology, Writing - review & editing. **Hirofumi Obinata:** Methodology, Writing - review & editing. **Hitoshi Mori:** Methodology, Writing - review & editing. **Kaku Tamura:** Methodology, Writing - review & editing. **Peng An:** Methodology, Writing - review & editing. **Bradford J. Wood:** Methodology, Writing - review & editing. **Daguang Xu:** Methodology, Writing - review & editing, Supervision.

References

American College of Radiology,. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., van Beek, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.-Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Vande Casteele, A., Gupte, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P., 2011. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Med. Phys. 38 (2), 915–931.

Bai, H.X., Wang, R., Xiong, Z., Hsieh, B., Chang, K., Halsey, K., Tran, T.M.L., Choi, J.W., Wang, D.-C., Shi, L.-B., et al., 2020. Ai augmentation of radiologist performance in distinguishing covid-19 from pneumonia of other etiology on chest ct. Radiology 201491.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based

cardiac mr image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 253–260.
van Berlo, B., Saeed, A., Ozcelebi, T., 2020. Towards federated unsupervised representation learning. In: Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking, pp. 31–36.
Bernheim, A., Mei, X., Huang, M., Yang, Y., Fayad, Z.A., Zhang, N., Diao, K., Lin, B., Zhu, X., Li, K., Li, S., Shan, H., Jacobi, A., Chung, M., 2020. Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection. Radiology 295 (3), 200463.
Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C., 2019. Remixmatch: semi-supervised learning with distribution alignment and augmentation anchoring arXiv:1911.09785.
Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. Mixmatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems, pp. 5050–5060.
Brisimi, T.S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I.C., Shi, W., 2018. Federated learning of predictive models from federated electronic health records. Int. J. Med. Inform. 112, 59–67.
Chen, Y., Qin, X., Wang, J., Yu, C., Gao, W., 2020. Fedhealth: a federated transfer learning framework for wearable healthcare. IEEE Intell. Syst..
Cheplygina, V., de Bruijne, M., Pluim, J.P., 2018. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis arXiv:1804.06353.
Clara, 2020. NVIDIA Clara Train Application Framework.
Colombi, D., Bodini, F.C., Petrini, M., Maffi, G., Morelli, N., Milanese, G., Silva, M., Sverzellati, N., Michieletti, E., 2020. Well-aerated lung on admitting chest ct to predict adverse outcome in covid-19 pneumonia. Radiology 0 (0), 201433.
Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C., 2019. Semi–supervised brain lesion segmentation with an adapted mean teacher model. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 554–565.
DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout arXiv:1708.04552.
Dong, D., Tang, Z., Wang, S., Hui, H., Gong, L., Lu, Y., Xue, Z., Liao, H., Chen, F., Yang, F., et al., 2020. The role of imaging in the detection and management of covid-19: a review. IEEE Rev. Biomed. Eng..
Drosten, C., Günther, S., Preiser, W., van der Werf, S., Brodt, H.-R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R..A., Berger, A., Burguière, A.-M., Cinatl, J., Eickmann, M., Escriou, N., Grywna, K., Kramme, S., Manuguerra, J.-C., Müller, S., Rickerts, V., Stürmer, M., Vieth, S., Klenk, H.-D., Osterhaus, A..D., Schmitz, H., Doerr, H..W., 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N top N. Engl. J. Med. 348 (20), 1967–1976.
Fan, D., Zhou, T., Ji, G., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Inf-net: Automatic covid-19 lung infection segmentation from ct images. IEEE Transactions on Medical Imaging. 1–1
Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Inf-net: automatic covid-19 lung infection segmentation from ct scans arXiv:2004.14133.
Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations arXiv:1803.07728.
Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D.S., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y., Zhong, N.-s., 2020. Clinical characteristics of coronavirus disease 2019 in china. N top N. Engl. J. Med. 382 (18), 1708–1720.
Guha, N., Talwlkar, A., Smith, V., 2019. One-shot federated learning arXiv:1902.11175.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
Hong, S., Noh, H., Han, B., 2015. Decoupled deep neural network for semi-supervised semantic segmentation. In: Advances in neural information processing systems, pp. 1495–1503.
Huang, L., Han, R., Ai, T., Yu, P., Kang, H., Tao, Q., Xia, L., 2020. Serial quantitative chest ct assessment of covid-19: deep-learning approach. Radiology: Cardiothoracic Imaging 2 (2), e200075.
Jin, Y., Wei, X., Liu, Y., Yang, Q., 2020. A survey towards federated semi-supervised learning arXiv:2002.11545.
Kang, H., Xia, L., Yan, F., Wan, Z., Shi, F., Yuan, H., Jiang, H., Wu, D., Sui, H., Zhang, C., Shen, D., 2020. Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning. IEEE Transactions on Medical Imaging. 1–1
Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks arXiv:1609.02907.
Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., Cao, K., Liu, D., Wang, G., Xu, Q., Fang, X., Zhang, S., Xia, J., Xia, J., 2020. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology 0 (0), 200905.
Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., et al., 2020. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology 200905.
Li, T., Sahu, A.K., Talwalkar, A., Smith, V., 2020. Federated learning: challenges, methods, and future directions. IEEE Signal Process. Mag. 37 (3), 50–60.
Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., et al., 2019. Privacy-preserving federated brain tumour segmentation. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 133–141.
Li, X., Yu, L., Chen, H., Fu, C.-W., Heng, P.-A., 2018. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. BMVC.
Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A., 2020. Semi-supervised medical image classification with relation-driven self-ensembling model arXiv:2005.07377.
Liu, S., Georgescu, B., Xu, Z., Yoo, Y., Chabin, G., Chaganti, S., Grbic, S., Piat, S., Teixeira, B., Balachandran, A., et al., 2020. 3D tomographic pattern synthesis for enhancing the quantification of covid-19 arXiv:2005.01903.
Liu, S., Xu, D., Zhou, S.K., Pauly, O., Grbic, S., Mertelmeier, T., Wicklein, J., Jerebko, A., Cai, W., Comaniciu, D., 2018. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 851–858.
Luo, Y., Zhu, J., Li, M., Ren, Y., Zhang, B., 2018. Smooth neighbors on teacher graphs for semi-supervised learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8896–8905.
McMahan, H.B., Moore, E., Ramage, D., Hampson, S., et al., 2016. Communication-efficient learning of deep networks from decentralized data arXiv:1602.05629.
Mei, X., Lee, H.-C., Diao, K.-y., Huang, M., Lin, B., Liu, C., Xie, Z., Ma, Y., Robson, P.M., Chung, M., et al., 2020. Artificial intelligence–enabled rapid diagnosis of patients with covid-19. Nat. Med. 1–5.
Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.
Misra, I., Shrivastava, A., Hebert, M., 2015. Watch and learn: Semi-supervised learning for object detectors from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3593–3602.
Mlynarski, P., Delingette, H., Criminisi, A., Ayache, N., 2019. Deep learning with mixed supervision for brain tumor segmentation. J. Med. Imaging 6 (3), 034002.
Myronenko, A., 2019. 3d mri brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, Cham, pp. 311–320.
Nguyen, T.D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., Sadeghi, A.-R., 2019. Dïot: A federated self-learning anomaly detection system for iot. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). IEEE, pp. 756–767.
Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision. Springer, pp. 69–84.
Ouyang, X., Huo, J., Xia, L., Shan, F., Liu, J., Mo, Z., Yan, F., Ding, Z., Yang, Q., Song, B., Shi, F., Yuan, H., Wei, Y., Cao, X., Gao, Y., Wu, D., Wang, Q., Shen, D., 2020. Dual-sampling attention network for diagnosis of covid-19 from community acquired pneumonia. IEEE Transactions on Medical Imaging. 1–1
Papandreou, G., Chen, L.-C., Murphy, K.P., Yuille, A.L., 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision, pp. 1742–1750.
Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A., 2018. Deep co-training for semi–supervised image recognition. In: Proceedings of the european conference on computer vision (eccv), pp. 135–152.
Shah, M.P., Merchant, S., Awate, S.P., 2018. Ms-net: mixed-supervision fully-convolutional networks for full-resolution segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 379–387.
Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., Shi, Y., 2020. Lung infection quantification of covid-19 in ct images with deep learning arXiv:2003.04655.
Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S., 2018. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In: International MICCAI Brainlesion Workshop. Springer, pp. 92–104.
Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D., 2020. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Reviews in Biomedical Engineering. 1–1
Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D., 2020. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Rev. Biomed. Eng..
Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., 2020. Fixmatch: simplifying semi-supervised learning with consistency and confidence arXiv:2001.07685.
Tang, Y., Wang, J., Gao, B., Dellandréa, E., Gaizauskas, R., Chen, L., 2016. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2119–2128.
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, pp. 1195–1204.
Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D., 2019. Interpolation consistency training for semi-supervised learning arXiv:1903.03825.
Wang, J., Bao, Y., Wen, Y., Lu, H., Luo, H., Xiang, Y., Li, X., Liu, C., Qian, D., 2020. Prior-attention residual learning for more discriminative covid-19 screening in ct images. IEEE Trans. Med. Imaging.

Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., Zheng, C., 2020. A weakly-supervised framework for covid-19 classification and lesion localization from chest ct. IEEE Transactions on Medical Imaging. 1–1

Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In: The IEEE Winter Conference on Applications of Computer Vision, pp. 3646–3655.

Xie, Q., Hovy, E., Luong, M.-T., Le, Q.V., 2019. Self-training with noisy student improves imagenet classification arXiv:1911.04252.

Xie, W., Jacobs, C., Charbonnier, J.-P., van Ginneken, B., 2020. Contextual two-stage u-nets for robust pulmonary lobe segmentation in ct scans of covid-19 and copd patients arXiv:2004.07443.

Xu, J., Wang, F., 2019. Federated learning for healthcare informatics arXiv:1911.06270.

Yang, Q., Liu, Y., Chen, T., Tong, Y., 2019. Federated machine learning: concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2), 1–19.

Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 605–613.

Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31 (3), 1116–1128.

Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L., 2019. S4l: Self-supervised semi-supervised learning. In: Proceedings of the IEEE international conference on computer vision, pp. 1476–1485.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., et al., 2020. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. Cell.

Zhou, T., Canu, S., Ruan, S., 2020. An automatic covid-19 ct segmentation based on u-net with attention mechanism arXiv:2004.06673.

Zhou, Y., Wang, Y., Tang, P., Shen, W., Fishman, E.K., Yuille, A.L., 2019. Semi-supervised multi-organ segmentation via multi-planar co-training. WACV.