electronics

MDPI

Article

Federated Learning for Medical Imaging Segmentation via Dynamic Aggregation on Non-IID Data Silos

Liuyan Yang [1,2], Juanjuan He [1,2,*], Yue Fu [1,2] and Zilin Luo [1,2]

[1] College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China
[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-Time Industrial System, Wuhan 430081, China
* Correspondence: hejuanjuan@wust.edu.cn

Abstract: A large number of mobile devices, smart wearable devices, and medical and health sensors continue to generate massive amounts of data, making edge devices' data explode and making it possible to implement data-driven artificial intelligence. However, the "data silos" and other issues still exist and need to be solved. Fortunately, federated learning (FL) can deal with "data silos" in the medical field, facilitating collaborative learning across multiple institutions without sharing local data and avoiding user concerns about data privacy. However, it encounters two main challenges in the medical field. One is statistical heterogeneity, also known as non-IID (non-independent and identically distributed) data, i.e., data being non-IID between clients, which leads to model drift. The second is limited labeling because labels are hard to obtain due to the high cost and expertise requirement. Most existing federated learning algorithms only allow for supervised training settings. In this work, we proposed a novel federated learning framework, MixFedGAN, to tackle the above issues in federated networks with dynamic aggregation and knowledge distillation. A dynamic aggregation scheme was designed to reduce the impact of current low-performing clients and improve stability. Knowledge distillation was introduced into the local generator model with a new distillation regularization loss function to prevent essential parameters of the global generator model from significantly changing. In addition, we considered two scenarios under this framework: complete annotated data and limited labeled data. An experimental analysis on four heterogeneous COVID-19 infection segmentation datasets and three heterogeneous prostate MRI segmentation datasets verified the effectiveness of the proposed federated learning method.

Keywords: federated learning; dynamic aggregation; knowledge distillation; COVID-19

Citation: Yang, L.; He, J.; Fu, Y.; Luo, Z. Federated Learning for Medical Imaging Segmentation via Dynamic Aggregation on Non-IID Data Silos. *Electronics* **2023**, *12*, 1687. https://doi.org/10.3390/electronics12071687

Academic Editors: Dawid Połap, Robertas Damasevicius and Hafiz Tayyab Rauf

Received: 1 March 2023
Revised: 30 March 2023
Accepted: 31 March 2023
Published: 3 April 2023

1. Introduction

Deep learning models have recently achieved remarkable progress in the medical image field, especially in medical image segmentation tasks [1–4]. However, traditional centralized training requires uploading all of the original data to the server to train a high-performance model. Under the framework of realistic privacy protection, data access limits within institutions and inter-agency data access barriers have resulted in the status quo of "data silos" [5]. Therefore, how to legally use healthcare data is a pressing problem in protecting data privacy.

Federated learning is a potential solution for handling "data silos" in the medical field [6]. Federated learning technology is one of the private computing branch technologies. Adhering to the principle that data are available and invisible, it can use data scattered among participants for joint analysis and modeling without uploading data to the server. Specifically, each medical institution/client trains a local model using its private data. It uploads the local model parameters (usually the gradients or weights) rather than the raw data to the server. Then, the server calculates the new global model by a weighted

aggregation of the locally trained models and distributes it back to the clients. Therefore, FL can obtain a model with a greater generalization ability without sharing the local datasets. As a result, it has recently attracted substantial attention in medical image diagnosis. Qayyum et al. [7] allowed multiple-edge medical institutions to use a federated learning framework to detect chest X-ray image abnormalities associated with COVID-19. Dou Qi et al. [8] present a federated learning method allowing multiple multinational institutions to detect COVID-19 lung abnormalities in CT images and externally validate patients from cross-border studies. Sarma et al. [9] verified the effectiveness of federated learning on three private prostate datasets, where the FL model has a better generalization performance than a single model.

Although these prior works have shown success for federated learning, they present new challenges. One of the critical challenges is statistical heterogeneity [10]. Since the local data on each client do not sample from the global joint distribution of all clients, local data cannot represent the overall global distribution. An interplay exists between local models, leading to model drift. Data heterogeneity in the medical field mainly includes quantity skew, feature distribution skew, and imaging acquisition skew. Quantity skew is where the data number of each hospital institution may vary greatly, and large hospitals often have more patients than small hospitals. Most medical images are gray-scale images with fuzzy borders and noise. Even with the same label, the feature distribution varies from customer to customer. Image acquisition skew refers to the difference in image quality caused by equipment and instruments used in different hospitals. Many FL methods have been shown to significantly reduce the performance where the data are non-IID [11–13]. Some works incorporated a proximal term in local optimization or changed the model aggregation scheme on the server side to cope with this issue. For example, FedProx [10] introduces an approximation term in the part of FedAvg to limit the size of local updates. FedBN [14] leaves the client's BN layer updated locally, and the aggregation of servers does not require an average BN layer. Some existing FL methods perform well using shallow neural network models on classification datasets. However, they have yet to be widely verified on deep neural networks, especially in segmentation tasks [15].

Another challenge with FL in the medical field is that it is rare for each medical institution/client to have rich labeled data. In the early stage of the COVID-19 epidemic, medical institutions needed more labeled data to train a high-precision model. The semi-supervised setting could use unlabeled data to solve the label-expensive problem, which is valuable in COVID-19 pandemic diseases. Recently, federated semi-supervised learning approaches have been proposed to solve this challenge. Liu et al. [16] proposed an inter-client relation matching scheme validated on brain CT and skin lesion classification datasets. Wu et al. [17] offered a federated contrastive learning framework on multi-institution cardiac MRI for volumetric medical image segmentation with limited annotations. However, these semi-supervised federated learning approaches require clients to share some extra parameters as supplementary information, which may leak privacy-sensitive information [18].

Federated semi-supervised learning includes two common scenarios, as shown in Figure 1. The first scenario (a) considers when both labeled and unlabeled data are on each client and the server contains no data. The second scenario (b) considers when the labeled data are only available at the server, and the clients have completely unlabeled data. We considered scenario (a) because it is closer to the actual situation. Medical institutions may not provide labeled data to the server, but it is realistic for the institutions themselves to have some less-labeled data.

This paper proposes a new framework, MixFedGAN, to solve the problem of medical image segmentation under non-IID data silos. In MixFedGAN, we developed a scheme to dynamically aggregate the global model according to the accuracy of the current client models and their differences so that outstanding customers can make more contributions to the global model. Moreover, we introduced knowledge distillation into the local generator model with a proposed new distillation regularization loss function. We also trained the global generator locally to protect the critical parameters of the global model from being

changed arbitrarily. The global generator is regarded as the teacher because the global generator has a stronger generalization ability and richer knowledge than the local models. While imitating the teacher's predictions, the local generator model's generalization ability can be improved to avoid over-fitting local data. An experimental analysis on four heterogeneous COVID-19 infection segmentation datasets and three heterogeneous prostate MRI datasets shows that our method can improve the model's convergence speed and generalization ability.

Additionally, we considered how the proposed framework can be used for the federated semi-supervised learning (FSSL) of the labels-at-client scenario. The experiments demonstrate that the proposed framework is still effective in handling semi-supervised scenarios. We also show that, when consistency regularization is used, the model's generalization ability in the semi-supervised scene can improve further.

The main contributions of this paper are summarized as follows:

- A new dynamic aggregation scheme was designed to enhance the stability and quality of segmentation results. We dynamically adjusted the client's weight during training instead of fixing a constant weight according to the number of samples or the average distribution.
- We proposed a distillation regularization loss function, i.e., using Kullback–Leibler divergence to guide the local generator model. It prevents essential parameters of the global generator model from significantly changing while tuning the global generator model on client-side local data.
- We considered the effectiveness of our framework in both supervised and semi-supervised scenarios and conducted an experimental analysis on four heterogeneous COVID-19 segmentation datasets and three heterogeneous prostate MRI datasets. Both comparative and ablation experiments show that our method is more stable and efficient.

The remainder of the paper is organized as follows: Section 2 gives a brief introduction to the relevant works in the literature; Section 3 presents the proposed framework and related schemes; Section 4 reports the result of our experiments; Section 5 provides some experimental discussions; Section 6 concludes.

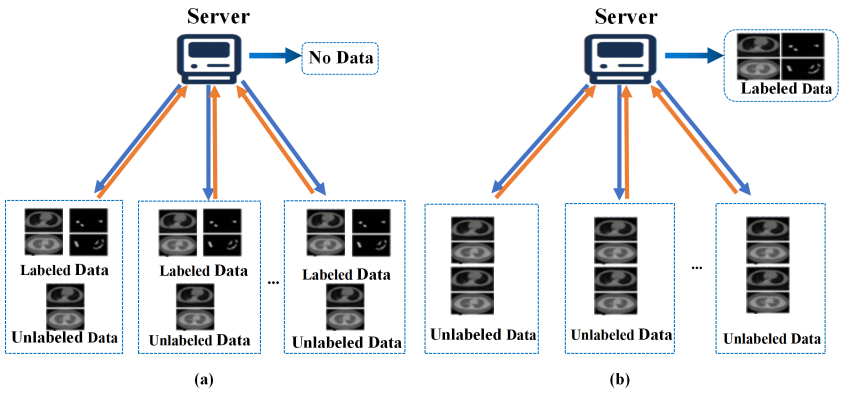

Figure 1. Illustration of two scenarios in federated semi-supervised learning. (**a**) Labels-at-Client Scenario: both labeled and unlabeled data are on each client. (**b**) Labels-at-Server Scenario: labeled instances are available on the server, while the clients have no labeled data.

2. Related Works and Motivation

The goal of federated learning is to achieve joint modeling to ensure data privacy, security, and legal compliance. It has played an essential role in medical image processing. Many FI methods combined with GAN can receive good results. In this section, we first briefly introduce the research progress of federated learning and its application in medical image processing. Then, the concept of GANs and some works integrated with federated learning are summarized. Finally, the research motivation of this paper is given.

2.1. Federated Learning

Federated learning (FL) was first proposed by McMahan et al. [19] in 2016 for predicting text input stored on tens of thousands of local Android machines. Some classic FL methods and their improvements have been widely verified in image classification. MOON [20] is designed as a simple and efficient method based on FedAvg [19]. It uses contrastive learning to add a new regularization function to constrain local models. FedNova [21] uses normalized averaging methods to eliminate target inconsistency while preserving fast error convergence. Scaffold [11] computes and aggregates control variates to correct the local updates. FedDyn [22] aligns the client using dynamic regularization to solve the client drift problem. Recent research on introducing data augmentation into federated learning has achieved promising results. FedMix [23] performs a mixup operation on average data from other clients and local data to approximate the global mixup. FedFA [24] allows the client to extract new samples for training from the universal statistic characterized by all participants to mitigate the client's feature drift.

Moreover, FL has achieved good application value in medical image segmentation. FL has been successfully applied on multi-institution brain MRI for tumor segmentation and the protection of patient information [25]. Lo et al. [26] evaluated the performance of a federated learning framework based on deep neural networks for retinal microvessel segmentation. Vaid et al. [27] allowed multiple institutions to use a federated learning framework that avoids collecting local data to predict mortality in hospitalized patients with COVID-19. We focus on medical image segmentation in the FL settings and FL semi-supervised settings.

2.2. Generative Adversarial Network

Generative adversarial networks (GANs) have received much attention for semantic segmentation and are also active in medical image segmentation. Some works have improved the generator and discriminator by combining the features of semantic segmentation and GAN. For example, Luc et al. [28] proposed a GAN for segmentation, where the generator tries to generate a segmentation map close to the ground. At the same time, the discriminator is used to distinguish between the two. Xue et al. [29] improved the previous work of Luc et al. and proposed a segmentation GAN (SegAN) where they used the critic to discriminate the multi-scale L_1 loss function instead of the value result (1 or 0) for giving more gradient feedback to the critic. Lei et al. [30] proposed a split with dual-discriminator network, trained with two supervised discriminator-assisted splits. A discriminator takes care of semantic context text inspection, and another is used for texture details examination. These methods achieve excellent segmentation accuracy in skin lesion segmentation.

Recently, the combined application of FL and GAN has been investigated. For example, Nguyen et al. [31] proposed a COVID-19 detection scheme that achieves privacy preserving and highly efficient COVID-19 detection by realizing the joint design of GAN and FL across medical institutions in edge–cloud computing. Rasouli et al. [32] proposed a federated generative confrontation network, FedGAN. It periodically synchronizes and broadcasts the parameters of the generator and discriminator through the intermediary and provides theoretical research on the convergence of FedGAN. Fan et al. [33] proposed four strategies for synchronizing local and central models. Zhang et al. [34] simulated a centralized discriminator by aggregating discriminators from all clients to learn the data distribution of different clients.

2.3. Motivation

Even though significant progress has been made on state-of-the-art designs, existing FL methods still need to overcome the following challenges, especially in medical image processing.

First, many medical image segmentation methods based on a federated learning framework follow the conventional settings, such as FedAvg, i.e., the client trains the model and sends the weights to the server, and the server performs simple average weighted

aggregation to obtain the final global model. In clinical practice, the quality and quantity of data from different clients may vary and not be independently and identically distributed. Some clients with a lower performance delay the convergence speed and reduce the model accuracy.

Second, most FL methods only consider image segmentation in supervised scenarios, based on client-side improvement or server-side optimization, with limited guidance, that cannot adapt to federated semi-supervised scenarios. These limitations motivate us to enhance the entire training process of FL locally and globally and use this method for FSSL on client-side labeling scenarios. Existing work on FSSL for unlabeled image segmentation in medical image analysis is minimal. Recently, Yang et al. [35] demonstrated the applicability of semi-supervised learning to COVID-19 pathological segmentation in a federated setting. However, Yang adopted a federated learning algorithm with a supervised and unsupervised client, which differs from the federated semi-supervised learning scenario.

3. Methods

We explored the problem of COVID-19 infection segmentation in a federal setting, considering four institutions collaborating by sharing network model parameters. Our goal is to improve the local and server models and reduce the impact of non-IID data. We proposed a new framework with an improved aggregation mechanism and introduced knowledge distillation in the local generator model with a new distillation regularization loss function. The generalization ability and stability of the global model were improved. This section presents our framework and its components.

3.1. Problem-Setting

Suppose that K institutions participate in federated learning. We denote (X,Y) as the joint image dataset and label space over K clients. Each client sample is an image–label pair (x,y) with $x \in X$, $y \in Y$ and denotes the dataset on the kth client as D_k. The x is the feature extracted from image x, and the joint probability P(x, y) can be written as P(y | x)P(x) and P(x | y)P(y). $L = \{l_k\}_{k=1}^{K}$ denotes a set of local models for K clients. Unlike the setting of IID, the dataset size of the K clients for joint learning is different. When the distribution of feature P(x) varies from different clients, the distribution of P(y | x) is the same. Sometimes, the distribution of P(x | y) is different from clients, and the P(y) is the same.

In the standard FL algorithm [19], the clients periodically update their local models with a stochastic gradient descent (SGD) optimizer and the learning rate η ($0 < \eta < 1$). The server collects C models' ($C \leq K$) local update parameter through the model averaging algorithm and updates global model ω in the next round as $\omega_{t+1} \leftarrow \omega_t - \eta \nabla l_k(\omega_k^t)$. Our method improves local models and aggregations to mitigate client drift and avoid accuracy degradation caused by non-IID data.

3.2. MixFedGAN-Supervised

The overview of the proposed framework MixFedGAN is described in Figure 2, which overcomes model drift caused by non-IID client data with a more stable performance. Our framework has a server and K clients with their local datasets. Each client obtains the parameters from the server for training the local GAN model. The rich knowledge of the global generator could help to reduce the bias of the local generator. Then, the server collects the model parameters of each client's generator and discriminator and aggregates them according to the model accuracy and model difference. The training is iterated over a certain period of communication rounds T until the desired accuracy performance is achieved.

Initially, each institution joints the training by updating the parameters of the generator (S stands for generator) and the critic (C stands for the critic) in each communication round (indexed by t). To make the output of the local model more reliable, we adopted a supervised adversarial training process based on multi-scale feature loss [29]. We multiplied

the local model predictions and ground truth separately with the input image to obtain the segmentation area and the actual lesion area of COVID-19. Next, we input two regions to the discriminator individually and obtained the corresponding features. The final loss function of the generator is:

$$L_{seg} = \sum_{h,w} \ell_{mae}(C(x_l \otimes S(x_l)), C(x_l \otimes y_l)) + \lambda_{dice}\ell_{dice}(S(x_l), y_l), \quad (1)$$

where h is the image height and w is the image width. ℓ_{mae} is the mean absolute error (MAE) or L_1 distance. $x_l \otimes S(x_l)$ is the multiplication of the input image and the generator's predicted label map. $x_l \otimes y_l$ is the multiplication of the input image and the ground truth label map. λ_{dice} is the dice loss coefficient, and the value is 0.1.

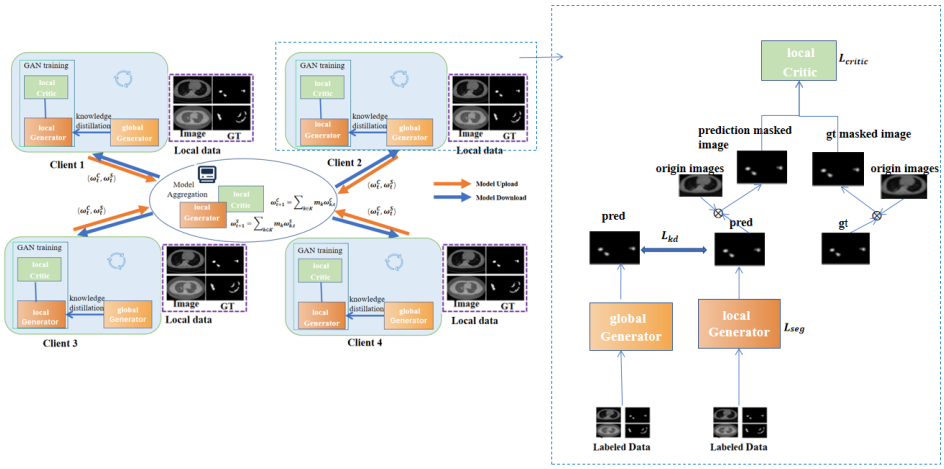

Figure 2. An overview of the privacy-preserving generative adversarial network framework (MixFedGAN) that deals with COVID-19 CT data from four regions using federated learning.

Because it is easy to explode or vanish the gradient of GAN, GAN has difficulties with actual training. We adopted the suggestion of Zhu et al. [3] and introduced a gradient penalty into the critic loss, which penalizes the gradient norm input of the discriminator. The final loss function of the discriminator is:

$$L_{critic} = \sum_{h,w} \ell_{mae}(C(x_l \otimes S(x_l)), C(x_l \otimes y_l)) + \lambda_d \mathbb{E}_{\tilde{x}\sim P_{\tilde{x}}}\left[(\|\nabla_{\tilde{x}} C(\tilde{x})\|_2 - 1)^2\right], \quad (2)$$

where λ_d is the gradient penalty coefficient, and the value is 0.0001. $P_{\tilde{x}}$ is uniform distribution along the straight lines between the pairs of points sampled from the actual and generator distribution [36].

Then, we exploited the rich knowledge of the global generator to reduce the bias of the local generator. Based on considering the accuracy of the current client model and the difference from the previous round of global models, the weighted dynamic aggregation mechanism was designed:

$$\begin{aligned} \omega^S_{t+1} &\leftarrow \sum_{k \in K} m_k \omega^S_{k,t} \\ \omega^C_{t+1} &\leftarrow \sum_{k \in K} m_k \omega^C_{k,t}, \end{aligned} \quad (3)$$

where m_k is the respective weight coefficient for each client. $\omega^C_{k,t}$ and $\omega^S_{k,t}$ are the model parameters of the generator and critic in the tth round for the kth client.

The formula for the supervised federated optimization objective is:

$$\min_{\omega^S \omega^C} \max l\left(\omega^S \omega^C\right) = \min_{\omega^S \omega^C} \max \left(L_{seg} + L_{kd} + L_{critic}\right), \tag{4}$$

In the supervised training process, the generator S aims to minimize the L_{seg} and L_{kd}. Critic C aims to maximize the ℓ_{mae} and minimize the gradient penalty loss in L_{critic}. Algorithm 1 shows the flow of our process. In Section 4, we demonstrate the advantages of MixFedGAN experimentally.

Algorithm 1 Training procedure of the proposed MixFedGAN

Input: Total number of clients K, generator S, critic C, total communication rounds T, local epochs E, learning rate η, B is mini-batchsize of data, if semi-supervised B_l, B_u is the local mini-batch size for labeled and unlabeled data. Initialization of a copy of global generator model weight $\omega^S_{k,t}$ and critic model weight $\omega^C_{k,t}$, the global segmentor model weight ω^T_t.

Output: Output the global model ω^S_{t+1}, ω^C_{t+1}

for each global round $t = 1, 2, \ldots, T$ **do**
  for each institution $k \in K$ **do**
    $\omega^S_{k,t}, \omega^C_{k,t} \leftarrow$ **LocalUpdate** $(k, \omega^S_t, \omega^C_t, \omega^T_t)$
  end for
  $\omega^S_{t+1}, \omega^C_{t+1} \leftarrow$ **ServerUpdate** $(k, \omega^S_{k,t}, \omega^C_{k,t})$
end for
LocalUpdate $(k, \omega^S_t, \omega^C_t, \omega^T_t)$: //Training in local
for local epoch e from 0 to *E-1* **do**
  if supervised **then**
    Sample minibatch B from D_k;
    Calculate the loss function according to (4)
  end if
  if semi-supervised **then**
    Sample minibatch B_l from D^k_l;
    Sample minibatch B_u from D^k_u;
    Calculate the loss function according to (9)
  end if
end for
send $\omega^S_{k,t}, \omega^C_{k,t}$ to the server
ServerUpdate ($k, \omega^S_{k,t}, \omega^C_{k,t}$): //Training in server
Calculate weight m_k for aggregation according to (5)
$\omega^S_{t+1} \leftarrow \sum_{k \in K} m_k \, \omega^S_{k,t}$, $\omega^C_{t+1} \leftarrow \sum_{k \in K} m_k \omega^C_{k,t}$
send $\omega^S_{t+1}, \omega^C_{t+1}$ to the clients

3.3. Dynamic Aggregation

When the raw data among clients taking part in federated learning are not independent and identically distributed, the model is prone to drift, and the model accuracy cannot be guaranteed. To tackle this issue, we designed a client dynamic aggregation weighting mechanism. It is no longer limited to the fixed index of the number of local data sets to determine the aggregation weight. However, it dynamically sets the weight for the client according to the current training situation and training information, achieving a more reasonable aggregation weight distribution.

Specifically, we first considered the test accuracy of each local model in each round. The aim is to make excellent local models significantly impact the aggregation of the global model, improving the quality of the global model. We denote the local model test accuracy as acc_k. Next, we added to account for the parameter difference between the currently trained local model and the previous round of the global model. We measured the L_2 distance between all ω^t_k and ω^t, where the former is the local model parameters in client k

and the latter is the global model parameters in each round t. The weight coefficient for each client m_k is:

$$m_k = \alpha \frac{1}{z_1} acc_k + \beta \frac{1}{z_2} \left\| \omega_k^t - \omega^t \right\|_2^2, \tag{5}$$

where $z_1 = \sum_{k \in K} acc_k$ and $z_2 = \sum_{k \in K} \left\| \omega_k^t - \omega^t \right\|_2^2$ are normalization factors. We normalized the calculated coefficients m_k again to bring them to the range of (0, 1) and $\sum_{k \in K} m_k = 1$. The coefficients of α and β were used to control the weight of these two terms. When the value of the l_2-norm is small, the behavior of the two parameter sets is more similar, or they are in the same direction.

3.4. Knowledge Distillation in MixFedGAN

Knowledge distillation (*KD*) is a kind of teacher–student training structure. Usually, the trained teacher model provides knowledge learning. The teacher's soft target provides significant regularization for the student model. The soft target imposes regularization training on the student model by providing label smoothing and a confidence penalty. The teacher network is fixed during the distillation, and only the student generator network is trained.

We used the previous round of the global generator model as the teacher and the local generator model as the student and let the teacher's rich knowledge learned from different data distributions serve as the student's supplementary knowledge. The global generator is regarded as the teacher because the global generator aggregates rich knowledge learned from different clients from the local data distribution and can represent the overall data distribution better than the local generator model. Kullback–Leibler (*KL*) divergence was used to reduce the difference between local generator network and global generator network models, and the distillation loss is as follows:

$$L_{kd} = KL(\sigma(T(x_i),\tau)||\sigma(S(x_i),\tau)), \tag{6}$$

where $T(x_i)$ and $S(x_i)$ denote the output of the global generator model T and local generator model S. τ is used to control smoothness in probabilities as the temperature term increases. It creates a more softened probability distribution. σ is the activation function. The soft targets of the teachers' model output carry more helpful information and improve the model generalization ability. We built a predictive map distillation module that enables the local generator network to learn prediction capabilities from the output feature map of the global generator network. We think of segmentation as a collection of pixel-level classification problems. We calculated the loss values for all pixel pairs at the same spatial location in both networks and combined these values into the distillation loss of this module.

3.5. MixFedGAN-Semi-Supervised

We applied MixFedGAN to the labels-at-client scenario in FSSL to address the challenge of limited labeling in federated learning. As shown in Figure 3, the proposed semi-supervised MixFedGAN consists of two loss terms: the labeled dataset $\{x_i^l, y_i^l\}_{i=1}^N$ calculates supervised losses L_{seg} and L_{critic}, and the unlabeled dataset $\{x_i^u\}_{i=1}^N$ calculates unsupervised losses $L_u^{distill}$ and L_u^{con}. The supervised function is defined as a multi-agent game of generators and critics, and has been discussed in Equations (1) and (2). For the unlabeled loss function, we focused on distillation using unlabeled samples x_u, and minimized the subsequent distillation loss:

$$L_u^{distill} = KL(\sigma(T(x_u),\tau)||\sigma(S(x_u),\tau)), \tag{7}$$

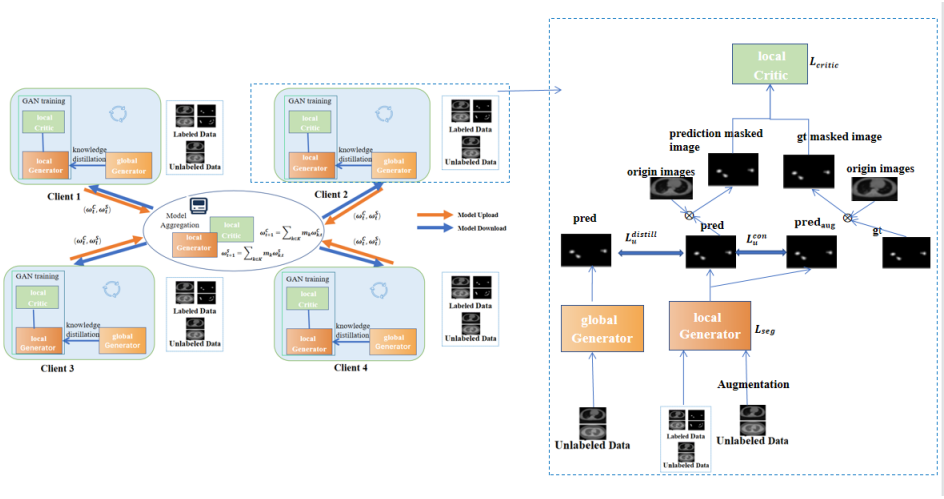

Figure 3. An overview of the privacy-preserving generative adversarial network framework (MixFedGAN) that deals with COVID-19 CT labeled and unlabeled data from four regions using federated learning.

We employed consistency regularization to learn from unlabeled data and improve the performance on labeled data. Consistency regularization [37] means that the prediction results of the model on the perturbed training samples should be consistent with the original prediction results. Since such methods do not rely on the true labeling of samples, large amounts of unlabeled data can be used. The consistency of the forecast is to hope that the two prediction results are as close as possible, i.e., the distance of D[p($(y \mid x)$,θ), $p(y \mid A(x),\theta)$] is as small as possible, where $D[p,q]$ is the distance measurement function, such as $|p-q|_2^2$ and $KL(KL\text{-}divergence)\sum p_i log\frac{p_i}{q_i}$. $A(\cdot)$ is a random data augmentation function, similar to random spatial translation, rotation, or adding noise. We aimed to minimize the distance between these two distributions using KL divergence. In $KL(p|q)$, p represents the real distribution and q represents the hypothetical distribution. However, in the consistency regularization method, p represents the predicted distribution of the original data and q represents the predicted distribution of the perturbed data. We denote the image augmentation of input x_u as $A(x_u)$. We represent the current predictions of samples as $S(y \mid x_u)$ and generated auxiliary predictions $S(y \mid A(x_u))$ from the perturbed samples $A(x_u)$. Therefore, the unsupervised consistency regularization loss for the kth client is defined as:

$$L_u^{con} = KL(S(y|x_u)||S(y|A(x_u))). \tag{8}$$

Finally, we updated the parameters according to the redesigned aggregation mechanism: $\omega_{t+1}^S \leftarrow \sum_{k \in K} m_k \omega_{k,t}^S$, $\omega_{t+1}^C \leftarrow \sum_{k \in K} m_k \omega_{k,t}^C$. The semi-supervised federated optimization objective formula is as follows:

$$\min_{\omega^S \omega^C} \max l\left(\omega^S \omega^C\right) = \min_{\omega^S \omega^C} \max\left(L_{seg} + \lambda_u L_u^{distill} + \lambda_{con} L_u^{con} + L_{critic}\right). \tag{9}$$

During generator and critic training with a min–max game in FSSL, the generator S aims to minimize the L_{seg}, $L_u^{distill}$, and L_u^{con} and the critic C aims to maximize the ℓ_{mae} and minimize the gradient penalty loss in L_{critic}. In our experiment, we set $\lambda_u = 1.0$, and $\lambda_{con} = 0.5$.

4. Experiments

This section evaluates our method on four publicly available COVID-19 CT scan segmentation datasets from different regions and three heterogeneous prostate MRI segmentation datasets. We conducted comparison experiments and an ablation study to verify the effectiveness of the proposed method and each component of our framework. The extensive evaluation demonstrates that our approach can improve local and global mod-

els without compromising client performance and achieves a higher accuracy and faster convergence than current state-of-the-art methods on non-IID datasets.

4.1. Dataset

To conveniently compare with others' work, we used the following four open-source COVID-19 segmentation datasets. Each dataset represents one FL client. The COVID-19-CT dataset is denoted as Client A. MS COVID-19 dataset is denoted as Client B. COVID-19-9 dataset is denoted as Client C. COVID-19-1110 dataset is denoted as Client D. These four datasets have different numbers of images and vary in size, shape, texture, and imaging protocols, as shown in Figure 4. The lesion areas of Client B and Client C are more apparent and prominent than others.

COVID-19-CT dataset [38]. The COVID-19 CT dataset was collected by Ma et al. [39], and comprises 20 annotated COVID-19 chest CT volumes. The volumes of each CT scan dataset subject have a resolution of 512 × 512 with slices of approximately 176 by mean (200 by median).

COVID-19-1110 dataset [40]. The COVID-19-1100 scan dataset comprises 1100 lung CT images of COVID-19 patients. The dataset was provided by medical hospitals in Moscow, Russia. Fifty of them were annotated, and ground-glass opacities (GGO) and consolidation regions were marked for lesion region segmentation.

COVID-19-9 dataset [41]. The COVID-19-9 dataset comprises nine axial COVID-19 volumetric CTs from Radiopaedia. A radiologist evaluated 373 out of the total of 829 slices as positive and labels including lungs and infected areas are present.

MS COVID-19-CT dataset [42]. The MS COVID-19 dataset was collected by the Italian Society of Medical and Interventional Radiology, and consists of 100 axial CT images from over 40 patients. The CT images were segmented by a radiologist using three labels: ground-glass opacity (GGO), consolidation, and pleural effusion.

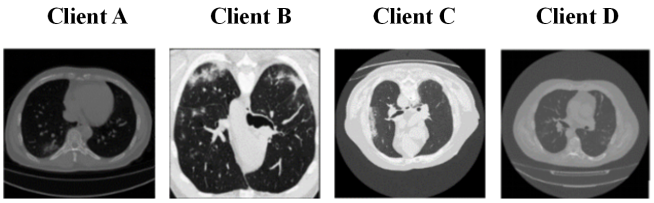

Figure 4. Examples of COVID-19 images from four clients showing considerable statistical heterogeneity.

We used the following three open-source prostate MRI segmentation datasets. We divided the data set into six parts, according to Liu et al. [43]. Each part represents one FL client. The BIDMC, HK, and UCL datasets are from the MICCAI 2012 Grand Challenge (PROMISE12). The I2CVB dataset is the Initiative for Collaborative Computer Vision Benchmarking (I2CVB). The BMC and RUNMC datasets are from the NCI-ISBI 2013 Challenge. We present a visualization of these datasets in Figure 5.

PROMISE12 dataset [44]. The PROMISE12 MRI dataset contains 50 training cases and 30 testing cases. These cases include a transversal T2-weighted MR image of the prostate. These scans were acquired using Siemens or GE scanners with or without endorectal coils and field strengths of 3 T or 1.5 T at different resolutions.

I2CVB dataset [45]. The I2CVB dataset is a multi-parameter dataset obtained from Siemens or GE scanners and field strengths of 3 T or 1.5 T at different resolutions.

NCI-ISBI 2013 dataset [46]. The NCI-ISBI 2013 dataset consists of 60 subjects. Among them, 30 cases are from the 1.5T scanner, and another 30 cases are from the 3 T scanner. The ground truth comprises four different classes: prostate, peripheral zone (PZ), central gland (CG), and cap.

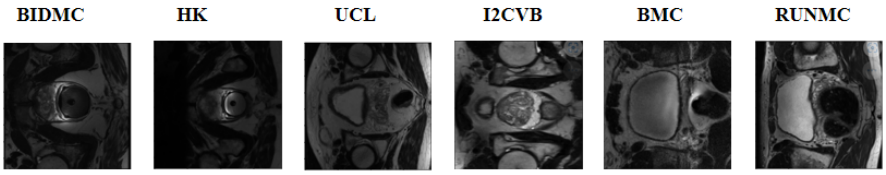

Figure 5. Examples of prostate MRI images from six clients showing considerable statistical heterogeneity.

4.2. Experimental Setup and Evaluation Metrics

Experimental Setup. We first demonstrated the effectiveness of our method in a supervised federated learning scenario, in which all clients have complete annotated data and ensure that no data are transferred among the clients except for centralized training. We compared it with current state-of-the-art FL methods, including FedAvg [19], FedProx [10], FedBN [14], MOON [20], and FedNova [21]. Moreover, for FedAvg, FedProx (μ = 0.001), and FedBN, we considered two different ways of assigning weights to clients. One is to allocate according to the number of samples that each client provides. We used "FedAvg", "FedProx", "FedBN" to represent. The other is that all clients have the same constant weight. We used "FedAvg-even", "FedProx-even", and "FedBN-even", respectively. The "centralized" model was trained on the aggregate training sample obtained from the clients' subsamples, which can be considered as an upper bound on FL. In the COVID-19 experiment, we compared local models that only use single-client data instead of aggregated updates, denoted by "Local only-A", "Local only-B", "Local only-C", and "Local only-D", respectively. In the prostate MRI experiment, we report the results of the FL methods running on the six clients. In addition, we include a comparison of the ablation study to verify the effectiveness of each component in our framework. "Proposed-1" represents the proposed method without knowledge distillation. "Proposed-2" describes the proposed process without dynamic aggregation.

In the COVID-19 training process, the batch size is 4, and we adopted the Adam optimizer with a learning rate of 2×10^{-4}, (β_1, β_2) of (0.5, 0.999), and learning rate decay of 0.5. The whole iteration number T is 100, and the number of local iterations defaults to 1. Each slice was resized to 256 $\times$ 256 as the input. For each client, we randomly split the annotated cases into training/testing data, resulting in splits of 1375/113 for Client A, 80/20 for Client B, 282/28 for Client C, and 588/58 for Client D. In the prostate MRI training process, the batch size is 4, and we adopted the Adam optimizer with a learning rate of 1×10^{-4}, (β_1, β_2) of (0.9, 0.999), and learning rate decay of 0.5. The whole iteration number T is 300, and the number of local iterations defaults to 1. Each slice was resized to 384 × 384 as the input. For each client, we randomly split the annotated cases into training/testing data, resulting in splits of 181/28 for the BIDMC dataset, 124/23 for the HK dataset, 143/27 for the UCL dataset, 331/85 for the I2CVB dataset, 347/92 for the BMC dataset, and 334/88 for the RUNMC dataset. All experiments were performed on a machine with GTX 3060 GPU and a 16 GB RAM equipped with PyTorch. All results reported are the average of three repeating runs.

Evaluation Metrics. We used three widely adopted metrics in the literature, including the Dice similarity coefficient (Dice), Sensitivity (Sen), and Accuracy (ACC). The formulae of Dice, Sen, and Acc are shown as follows:

$$Dice = \frac{2 \times TP}{FN + 2 \times TP + FP}, \tag{10}$$

$$Sen = \frac{TP}{FN + TP}, \tag{11}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{12}$$

where *TP*, *TN*, *FP*, and *FN* refer to true positive, true negative, false positive and false negative pixels of the output images and the ground truth.

4.3. Comparison Experiments and Ablation Study

The quantitative results are shown in Table 1. The first column (Client A) indicates the model's test results using only Client A's testing data. It can reflect the generalization performance of the model on Client A. Columns 2–4 are also similar. The fifth column (Avg.) is abbreviated for the average accuracy in terms of Dice. It is an important indicator used to measure the performance of the global model.

In Table 1, we first trained each client's local model using only their data, and the results are shown in the first four rows. The local model performs best on its own test set, and the generalization performance on other clients' testing data is much lower. Combining all datasets in centralized training can improve the model's accuracy and generalization, which can be seen as an upper bound.

Then, we compared the proposed method with five current state-of-the-art FL methods. All forms operate under the same conditions, i.e., the global model can collect information (the gradients and weights) from multiple local models but cannot obtain client data for privacy reasons. We can find that the proposed method outperformed other methods and is closer to or even better than the centralized one. We attributed this improvement to knowledge distillation and dynamic aggregation, enhancing the training process on both the client and server sides. When all clients have the same constant weight, "FedAvg-even", "FedProx-even", "MOON-even',' and "FedNova" achieve similar results on "Avg.". "FedBN-even" outperforms "FedAvg-even" by 0.61% on "Avg." but only improves the accuracy of clients A, B, and C in terms of Dice compared to "FedAvg-even". These methods do not achieve results close to centralized training, as improvements are limited only from the client or server sides. Our proposed method outperforms "FedAvg-even" by 3.4% on "Avg." and improves the accuracy of clients A, B, C, and D compared to "FedAvg-even" by 4.13%, 5.37%, 0.92%, and 3.17% in terms of Dice. With client-side improvement and server-side optimization, our method consistently outperforms others, reaching an accuracy of 72.96% on "Avg.", which is 2.79% higher than the previous state-of-the-art method ("FedBN-even") for the non-IID dataset.

These federated learning methods show limited progress when assigning weight according to the number of samples that each client provides. This is because, when the sample sizes of clients vary widely, clients with large sample sizes will dominate when the model is aggregated, causing the global network to shift towards the clients with large sample sizes and obliterating the contribution of clients with small sample sizes.

Compared with "FedAvg-even", proposed-1 adopts dynamic aggregation, improving the accuracy of clients A, B, and C by 2.25%, 4.42%, and 0.66% in terms of Dice, outperforming "FedAvg-even" by 1.67% on "Avg.". Proposed-2 adopts knowledge distillation, improving the accuracy of clients A, B, C, and D by 2.83%, 1.89%, 0.4%, and 1.88% in terms of Dice, outperforming "FedAvg-even" by 1.75% on "Avg.". Although proposed-1 did not improve the accuracy of client D compared to "FedAvg-even", the reason may be that the local test accuracy of client D is too low and the weight assigned is small. Still, the purpose of proposed-1 is to allow clients with a high local test accuracy to gain more weight. Proposed-2 transfers the global generator model information to local generator training, improving the accuracy of all clients compared to "FedAvg-even". In summary, our proposed combinations make sense because each component individually beats "FedAvg-even" in terms of "Avg.".

In addition, we visualized the segmentation results to demonstrate a qualitative comparison. As shown in Figure 6, all of these methods' results were roughly the same in terms of segmentation boundaries. Compared with the ground truth, other FL methods over-segment or miss some of the COVID-19 lesion areas. The proposed method can segment the infected area of COVID-19 with more details than other FL methods.

Table 1. Multi-national COVID-19 lesion segmentation. T is the total time required for model training in minutes.

Method	Client A			Client B			Client C			Client D				
	Dice	Sen	Acc	Dice	Sen	Acc	Dice	Sen	Acc	Dice	Sen	Acc	Avg. (Dice)	T (min)
Local only-A	72.35	82.18	99.56	53.80	34.18	98.20	57.12	47.40	95.81	33.94	21.03	98.28	54.30	168
Local only-B	53.89	71.71	99.16	76.87	57.96	98.93	73.73	77.93	96.73	44.76	53.28	98.62	62.31	17
Local only-C	49.85	56.35	99.12	65.63	44.18	98.58	75.99	73.19	97.27	56.60	58.76	99.20	62.02	36
Local only-D	55.63	55.01	99.40	58.29	37.71	98.34	71.50	73.84	96.06	61.88	63.10	99.31	61.83	75
Centralized	72.18	79.73	99.58	81.15	59.10	99.16	80.40	80.91	97.68	59.12	58.97	99.28	73.21	225
FedAvg	71.67	71.27	99.59	64.53	42.54	98.57	52.71	35.79	95.88	42.02	33.79	99.17	57.73	230
FedProx	71.01	78.84	99.56	65.05	43.98	98.55	64.84	52.03	96.68	57.93	51.03	99.24	64.70	255
FedBN	72.23	79.06	99.58	64.30	43.23	98.53	63.84	56.02	96.26	41.78	33.10	99.18	60.54	223
FedAvg-even	70.52	78.84	99.55	74.59	52.59	98.90	76.87	72.99	97.44	56.26	53.10	99.29	69.56	230
FedProx-even	69.27	69.04	99.58	73.18	50.13	98.85	75.27	73.63	97.15	58.50	53.79	**99.33**	69.05	255
FedBN-even	72.17	84.63	99.55	76.18	55.03	98.92	76.96	69.77	97.54	55.38	51.90	99.26	70.17	223
MOON-even	68.07	78.17	99.37	76.01	55.17	98.91	77.16	69.93	97.56	56.42	55.86	99.27	69.41	385
FedNova	69.22	86.64	99.47	76.42	55.07	98.96	75.94	73.97	97.24	55.17	50.17	99.28	69.18	231
Proposed	**74.65**	**88.20**	**99.60**	**79.96**	**57.86**	**99.11**	**77.79**	74.75	**97.49**	**59.43**	**61.72**	99.26	**72.96**	273
Proposed-1	72.77	77.06	99.60	79.01	58.51	99.05	77.53	**76.35**	97.39	55.62	60.69	99.17	71.23	232
Proposed-2	73.35	79.06	99.61	76.48	55.72	98.95	77.27	73.53	97.45	58.14	56.03	99.29	71.31	272

The results of the prostate MRI dataset are reported in Table 2. We compared our method with "FedAvg", "FedProx", "FedBN", "MOON", and "FedNova" with the same constant weight. "FedAvg" and "FedNova" achieve similar results on "Avg.". "MOON" outperforms "FedAvg" by 0.22% on "Avg.". "FedBN" outperforms "FedAvg" by 0.68% on "Avg.". Our proposed method outperforms "FedAvg" by 1.72% on "Avg." and can improve all client's accuracy in terms of Dice compared to "FedAvg" by 2.09%, 1.53%, 1.42%, 1.15%, 2.26%, and 1.84%. Compared with "FedAvg", proposed-1 adopts dynamic aggregation, outperforming "FedAvg" by 0.97% on "Avg.". Proposed-2 adopts knowledge distillation, outperforming "FedAvg" by 1.08% on "Avg.". Although proposed-1 did not improve the accuracy of the BIDMC dataset compared to "FedAvg", the reason may be that the BIDMC dataset is blurry and the weight assigned is small. Still, the purpose of proposed-1 is to allow clients with a high local test accuracy to gain more weight. Proposed-2 transfers the global generator model information to local generator training, improving the accuracy of all clients compared to "FedAvg".

Table 2. Multi-national prostate MRI segmentation. Dice score (%) is reported. T is the total time required for model training in minutes.

Method	BIDMC	HK	UCL	I2CVB	BMC	RUNMC	Average	T (min)
FedAvg	88.38	92.33	90.78	90.71	90.42	93.70	91.05	846
FedProx	87.63	92.17	90.14	90.65	91.04	93.89	90.92	921
FedBN	88.16	93.37	91.60	91.42	91.72	94.13	91.73	834
MOON	88.56	92.85	90.95	90.56	91.13	93.58	91.27	1224
FedNova	87.24	92.73	90.64	90.92	90.63	93.97	91.02	849
Proposed	**90.47**	**93.86**	**92.20**	**91.86**	**92.68**	**95.54**	**92.77**	984
Proposed-1	88.24	93.69	92.15	91.65	92.16	94.24	92.02	852
Proposed-2	89.65	93.76	91.83	91.10	91.43	95.01	92.13	979

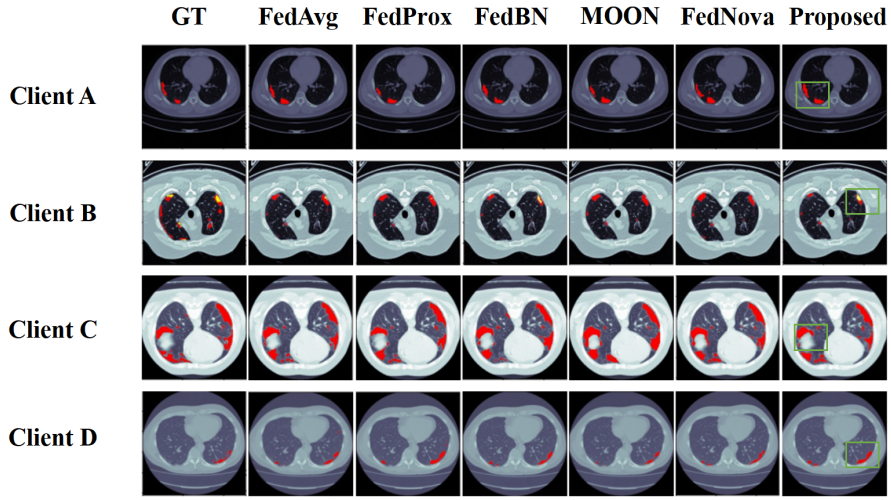

Figure 6. Visual comparison of COVID-19 infection regions segmentation results on the four clients, where the red labels denote COVID-19 infection regions, and the green boxes highlight some segmentation details.

4.4. Dealing with Semi-Supervised Setting

We considered another scenario with limited labeled data that was closer to reality. In previous works, some semi-supervised methods have achieved performance results in a traditional centralized training scenario. For example, mean teacher (MT) [47] used the EMA of the student model weights and calculated the MSE distance between teacher and student predictions. Virtual adversarial training (VAT) [48] used random perturbations to change the current model's predictions on unlabeled data substantially. Interpolation consistency training (ICT) [49] encouraged the predictions interpolated for unlabeled data to be consistent with the interpolation of predictions. Deep adversarial networks (DANs) [50] employ adversarial learning to distinguish between labeled and unlabeled data. We used "FedAvg" under the same constant weight setting with a naive combination of some semi-supervised methods to compare our experiments. We evaluated the model performance under the 10% and 20% labeled data settings using partly labeled federated data training as a baseline on FSSL and fully labeled data federated training as an upper bound on FSSL.

Table 3 presents the results of semi-supervised training with 10% and 20% labeled labels. As observed, both "FedAvg-MT" and "FedAvg-DNA" can improve the accuracy of clients A and C on the baseline but perform poorly on clients B and D. "FedAvg-ICT" improves the accuracy of clients A, B, and C on the baseline but perform poorly on client D. "FedAvg-VAT" improves the accuracy of clients A, B, and C on the baseline and outperforms the baseline by 1.37% on "Avg." under the 10% labeled scenario. It also exceeds the baseline by 1.61% on "Avg." under the 20% labeled scenario. However, the FedAvg naively combined with semi-supervised methods does not improve the accuracy of all clients, and all combinations perform poorly on client D. The reason may be that existing consistency-based semi-supervised methods are built with a single trainable model, which cannot provide information from multiple clients to enrich the unsupervised knowledge for unlabeled data in an FL setting.

Compared with these methods and the baseline, our method achieves a higher accuracy on all clients. The method based on dynamic aggregation (*DA*) and knowledge distillation (*KD*) can outperform the baseline by 2.41% on "Avg." under the 10% labeled scenario and exceed the baseline by 3.06% on "Avg." under the 20% labeled scenario. Based on the perturbation of unlabeled data by consistency regularization, we can further improve the model's generalization performance. It outperforms the baseline by 3.71% on

"Avg." under the 10% labeled scenario and the baseline by 4.33 % on "Avg." under the 20% labeled scenario. Our method is effective under FSSL because the local and global generator models can produce different decision boundaries for unlabeled data. The global generator model's viewpoint can be regarded as complementary knowledge of the local generator model.

Regarding the time cost, our MixFedGAN takes about the same time as the semi-supervised method mean teacher, but our accuracy is higher. If the consistency regularization method is added, the prediction will be further improved, but the training time will be longer due to the more complex loss function. In summary, our method is superior in terms of the time cost and accuracy compared to FedAvg naively combined with other semi-supervised methods.

Table 3. Comparison with a naive combination of some semi-supervised methods and FedAvg, using limited labeled data on multi-national COVID-19 lesion segmentation.

Method	Labeled	Unlabeled	Client A			Client B			Client C			Client D				
			Dice	Sen	Acc	Dice	Sen	Acc	Dice	Sen	Acc	Dice	Sen	Acc	Avg. (Dice)	T (min)
FedAvg-Fully	100%		70.52	82.18	99.52	74.59	52.59	98.90	76.87	72.99	97.44	56.88	53.10	99.29	69.72	230
FedAvg-Partly	10%		62.01	68.17	99.38	53.01	33.73	98.17	50.97	36.29	95.89	48.70	33.97	99.37	53.67	58
FedAvg-MT	10%	90%	63.01	69.27	99.36	45.29	28.61	97.88	51.31	38.38	95.71	39.37	28.62	99.04	49.75	268
FedAvg-DAN	10%	90%	62.40	68.72	99.39	49.26	32.78	97.80	53.76	41.34	95.81	45.22	40.34	99.13	52.66	318
FedAvg-VAT	10%	90%	64.84	70.69	99.45	52.87	33.18	98.17	54.77	42.29	95.88	47.67	37.93	99.26	55.04	276
FedBN-ICT	10%	90%	63.16	68.93	99.41	53.13	34.58	98.15	52.38	43.59	95.14	45.37	34.24	99.28	53.51	277
Proposed(KD+DA)	10%	90%	63.81	69.82	99.43	53.62	34.63	98.16	58.13	46.31	96.06	48.76	40.69	99.24	56.08	272
Proposed(KD+DA+Consistency)	10%	90%	**64.95**	**71.65**	**99.46**	**54.75**	**35.42**	**98.20**	**60.43**	**47.97**	**96.30**	**49.39**	**41.90**	**99.29**	57.38	330
FedAvg-Partly	20%		66.06	70.82	99.47	64.07	41.74	98.56	62.40	49.57	96.48	51.05	43.97	99.25	60.90	71
FedAvg-MT	20%	80%	68.35	78.40	99.50	54.96	33.93	98.29	62.95	50.48	96.50	43.63	38.17	99.11	57.47	236
FedAvg-DAN	20%	80%	67.59	73.72	99.53	55.06	34.78	98.26	64.03	51.98	96.56	46.02	38.41	99.14	58.18	285
FedAvg-VAT	20%	80%	69.47	78.15	99.59	64.17	42.64	98.34	67.22	57.89	96.67	49.19	42.07	99.23	62.51	243
FedAvg-ICT	20%	80%	66.46	71.05	99.51	64.67	44.45	98.43	63.61	52.08	96.49	47.84	38.28	99.26	60.65	244
Proposed(KD+DA)	20%	80%	68.75	74.61	99.54	64.84	**45.32**	98.49	70.28	64.28	96.67	51.98	39.66	**99.35**	63.96	238
Proposed(KD+DA+Consistency)	20%	80%	**70.41**	**78.85**	**99.61**	**66.03**	42.89	**98.65**	**71.69**	**65.08**	**97.02**	**52.80**	**45.38**	99.30	65.23	295

5. Discussion

5.1. Convergence Analysis

In this paper, our method was used for reducing local modal and global modal drifts, which partly improves learning efficiency. The Dice similarity coefficient (Dice) results using different communication rounds are plotted in Figure 7. In Figure 7a–c, our method increases smoothly with an increase in the number of communication rounds. In contrast, other FL methods show unstable convergence and a lower accuracy. In Figure 7d, the overfitting phenomenon inevitably occurred due to the single test sample of client D. However, it can be seen that, because our method alleviates a certain degree of overfitting, the prediction of our method on client D is also better than other FL methods. Therefore, the proposed method can achieve better results with fewer communication rounds, accelerate the convergence speed, and improve the model's generalization ability.

5.2. Influence of α and β

The aggregation mechanism that we designed calculates the contribution score for each client in each round according to the model accuracy index and the model difference index. Among them, α and β in (5) are used to distinguish the importance of the model accuracy index and the model difference index. Specifically, the significance of the model accuracy index is greater than that of the model difference index. The whole training is to dynamically set the weight for the client, which realizes a more reasonable aggregation weight distribution.

As we gradually increased the coefficient of β, the weights of those clients whose model parameters differ significantly from the global model parameters increased. In our experiments, the weight of client A varies the most from the global model, and the reason for this may be that the generalization performance of A is the worst. Increasing the β coefficient will cause the weight of client A to increase, thereby improving the accuracy of client A. It can be seen from Figure 8 that the accuracy of client A will reduce the accuracy

of the other three clients simultaneously. Since the samples of clients B and C are relatively small, but the images of clients B and C are fairly clear, the image quality is better, and the test results of their local models are also better. If aggregated according to the amount of sample data, the contributions of clients B and C will be obliterated by the client with large samples. Therefore, we aggregated them according to the model accuracy and model difference during training to give them a greater weight to speed up the training process. In this experiment, we set α to 0.8 and β to 0.2.

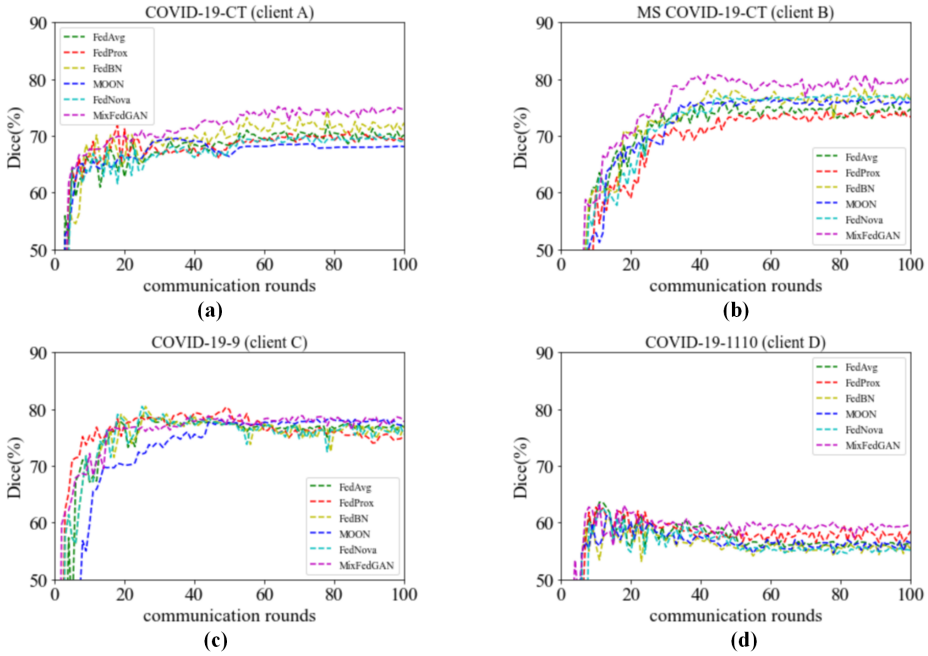

Figure 7. (**a**) Client A's convergence in terms of testing dice with communication rounds. (**b**) Client B's convergence in terms of testing dice with communication rounds. (**c**) Client C's convergence in terms of testing dice with communication rounds. (**d**) Client D's convergence in terms of testing dice with communication rounds.

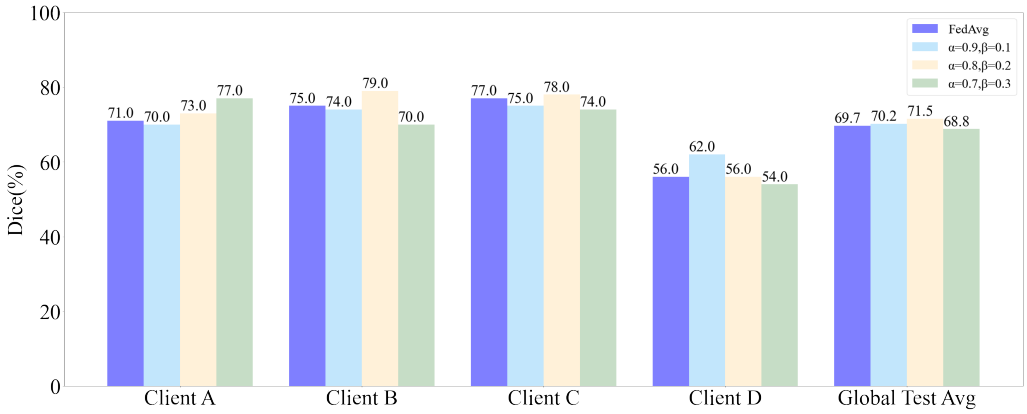

Figure 8. Visualizations of the effect of different α and β on the training results.

5.3. Influence of Distillation Temperature

Reference [51] showed that a soft temperature larger than one is critical for the effectiveness of *KD*. Increasing the temperature τ for the model output in (6) generates a smoother probability distribution. We gradually increased the distillation temperature, increasing the temperature from 1 to 20, and performed ablation experiments on the hyperparameter τ on proposed-2. As seen in Figure 9, with an increase in temperature, the accuracy of the proposed-2 method increases on the C and D clients, and the accuracy

of the A and B clients decreases. We can also see that each client's Dice score is better than the baseline "FedAvg" when the temperature equals 15, indicating that introducing global knowledge for distillation benefits client learning. In this experiment, we set τ to 15 because this temperature performs best.

Figure 9. Visualization of the effect of different distillation temperatures on the training results.

5.4. Computation and Communication Cost

High computational and communication costs have always been an urgent problem to be solved in federated learning. The total cost of FedAvg includes the computing cost of the local client and the upload/download communication cost between the client and the server. The improvement in the FedAvg algorithm inevitably introduces local computing costs, upload and download communication costs, and even server-side computing costs. Table 4 shows the average training time per round. FedProx introduces an approximation term in the local model, which nearly doubles the computational cost of local training. MOON presents the previous round of the global model and the local model in local training, which increases the computing resources of local training by nearly triple and significantly increases the training time. FedNova assigns client weights by normalizing the number of local iterations, and each client sends the normalized parameters update to the server, which adds a negligible communication cost to the server. FedBN ignores the exchange of the BN layer in the server aggregation part, and the computing resources that it introduces to the server are negligible.

Our method introduces the previous round's global generator model in the local client, inevitably increasing the local computation cost. In addition, regarding the communication cost, although the weights of client-side aggregation need to be recalculated on the server side, it will add negligible communication cost to the server because the client only needs to send the two indicators of local accuracy and model difference to the server for aggregating the model. Our total round training time increased compared to FedAvg. From the actual running time in Table 3, the time increase is within the acceptable range. However, our method can achieve a higher accuracy than FedAvg. We can still consider our approach as adequate. In future work, we will also solve the problem of the local computational cost through model compression and quantization.

Table 4. The average training time per round.

Method	COVID-19	Prostate MRI
FedAvg	2 min 16 s	2 min 49 s
FedProx	2 min 31 s	3 min 04 s
FedBN	2 min 15 s	2 min 48 s
MOON	3 min 51 s	4 min 05 s
FedNova	2 min 17 s	2 min 50 s
MixFedGAN	2 min 44 s	3 min 17 s

6. Conclusions

In this paper, we proposed a new framework MixFedGAN for automatic COVID-19 infection segmentation in order to mitigate client drift caused by non-IID data. The dynamic aggregation mechanism was designed to reduce the impact of current low-performing clients and improve stability. Knowledge distillation with a new distillation regularization loss function prevents essential parameters of the global generator model from significantly changing while tuning the global generator model on client-side local data. We also considered both supervised and semi-supervised scenarios to verify our methods. Four public COVID-19 CT scan datasets were employed for qualitative and quantitative analysis. The experiment shows that our proposed method can obtain high-quality segmentation results and outperforms some state-of-the-art FL methods. The model test in the supervised federal scenario is increased by 3.4% in terms of Dice compared with FedAvg in the COVID-19 dataset and increased by 1.72% in terms of Dice compared with FedAvg in the prostate MRI dataset. In the semi-supervised federated scenario, the model test is improved by 3.71% in terms of Dice compared with the baseline algorithm when the label is 10%. Since federated learning is in its infancy, problems still need to be solved, such as a limited communication bandwidth, limited mass data storage of edge node devices, and model parameters and gradients that are vulnerable to malicious attacks. To tackle the challenges of federated learning in communication efficiency and security, in future work, we plan to reduce the number of model parameters through a model compression technique to reduce the communication time of federated learning without sacrificing much accuracy and prevent the leakage of the original model.

Author Contributions: Conceptualization, L.Y.; methodology, L.Y.; software, L.Y.; validation, L.Y., Y.F. and Z.L.; formal analysis, L.Y., Y.F., and Z.L.; investigation, L.Y. and J.H.; resources, J.H.; writing—original draft preparation, L.Y.; writing—review and editing, L.Y. and J.H.; supervision, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China under Grant 62272355, 61702383, and 62176191.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. COVID-19-CT Database [38]: Official webpage link https://www.kaggle.com/andrewmvd/covid19-ct-scans/ (accessed on 9 June 2022). COVID19-1110 Database [40]: Official webpage link https://healthcaresummit.ieee.org/data-hackathon/ieee-covid-19-imaging-informatics-challenge/ (accessed on 2 February 2023). COVID-19-9 Database [41]: Official webpage link http://medicalsegmentation.com/covid19/ (accessed on 2 June 2022). MS COVID-19-CT dataset DataBase [42]: Official webpage link https://sirm.org/category/senza-categoria/covid-19/ (accessed on 2 June 2022). PROMISE12 Database [44]: Official webpage link Available: https://promise12.grand-challenge.org/ (accessed on 20 March 2023). I2CVB Database [45]: Official webpage link https://i2cvb.github.io/ (accessed on 20 March 2023). NCI-ISBI 2013 Database [46]: Official webpage link https://wiki.cancerimagingarchive.net/display/Public/NCI-ISBI+2013+Challenge+-+Automated+Segmentation+of+Prostate+Structures/ (accessed on 20 March 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou, T.; Li, L.; Bredell, G.; Li, J.; Unkelbach, J.; Konukoglu, E. Volumetric memory network for interactive medical image segmentation. *Med. Image. Anal.* **2023**, *83*, 102599. [CrossRef] [PubMed]
2. Liu, X.; Yuan, Q.; Gao, Y.; He, K.; Wang, S.; Tang, X.; Tang, J.; Shen, D. Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images. *Pattern Recognit.* **2022**, *122*, 108341. [CrossRef]
3. He, J.; Zhu, Q.; Zhang, K.; Yu, P.; Tang, J. An evolvable adversarial network with gradient penalty for COVID-19 infection segmentation. *Appl. Soft Comput.* **2021**, *113*, 107947. [CrossRef]
4. Liu, X.; Guo, Z.; Cao, J.; Tang, J. MDC-net: A new convolutional neural network for nucleus segmentation in histopathology images with distance maps and contour information. *Comput. Biol. Med.* **2021**, *135*, 104543. [CrossRef] [PubMed]

5. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. [CrossRef]
6. Xu, J.; Glicksberg, B.S.; Su, C.; Walker, P.; Bian, J.; Wang, F. Federated Learning for Healthcare Informatics. *J. Healthc. Inform. Res.* **2019**, *5*, 1–19. [CrossRef]
7. Qayyum, A.; Ahmad, K.; Ahsan, M.A.; Al-Fuqaha, A.; Qadir, J. Collaborative Federated Learning for Healthcare: Multi-Modal COVID-19 Diagnosis at the Edge. *IEEE Open J. Comput. Soc.* **2022**, *3*, 172–184. [CrossRef]
8. Dou, Q.; So, T.Y.; Jiang, M.; Liu, Q.; Vardhanabhuti, V.; Kaissis, G.; Li, Z.; Si, W.; Lee, H.H.C.; Yu, K.; et al. Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study. *NPJ Digit. Med.* **2021**, *4*, 1–11. [CrossRef]
9. Sarma, K.V.; Harmon, S.; Sanford, T.; Roth, H.R.; Xu, Z.; Tetreault, J.; Xu, D.; Flores, M.G.; Raman, A.G.; Kulkarni, R.; et al. Federated learning improves site performance in multicenter deep learning without data sharing. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1259–1264. [CrossRef]
10. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. In Proceedings of the Machine Learning and Systems (MLSys 2020), Austin, TX, USA, 2–4 March 2020; pp. 429–450. [CrossRef]
11. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. Scaffold: Stochastic controlled averaging for federated learning. In Proceedings of the 37th International Conference on Machine Learning (ICML 2020), Vienna, Austria, 13–18 July 2020; pp. 5132–5143. [CrossRef]
12. Li, X.; Huang, K.; Yang, W.; Wang, S.; Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv* **2019**, arXiv:1907.02189.
13. Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; Chandra, V. Federated learning with non-iid data. *arXiv* **2018**, arXiv:1806.00582.
14. Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; Dou, Q. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv* **2021**, arXiv:2102.07623.
15. Zhu, H.; Xu, J.; Liu, S.; Jin, Y. Federated learning on non-IID data: A survey. *Neurocomputing* **2021**, *465*, 371–390. [CrossRef]
16. Liu, Q.; Yang, H.; Dou, Q.; Heng, P.A. Federated semi-supervised medical image classification via inter-client relation matching. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention-MICCAI 2021, Strasbourg, France, 27 September–1 October 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 325–335. [CrossRef]
17. Wu, Y.; Zeng, D.; Wang, Z.; Shi, Y.; Hu, J. Federated contrastive learning for volumetric medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2021, Strasbourg, France, 27 September–1 October 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 367–377. [CrossRef]
18. Sun, J.; Li, A.; Wang, B.; Yang, H.; Li, H.; Chen, Y. Provable defense against privacy leakage in federated learning from representation perspective. *arXiv* **2020**, arXiv:2012.06043.
19. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282. [CrossRef]
20. Li, Q.; He, B.; Song, D. Model-contrastive federated learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10713–10722. [CrossRef]
21. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Adv. Neural. Inf. Process. Syst.* **2020**, *33*, 7611–7623. [CrossRef]
22. Acar, D.A.E.; Zhao, Y.; Navarro, R.M.; Mattina, M.; Whatmough, P.N.; Saligrama, V. Federated learning based on dynamic regularization. *arXiv* **2021**, arXiv:2111.04263.
23. Yoon, T.; Shin, S.; Hwang, S.J.; Yang, E. Fedmix: Approximation of mixup under mean augmented federated learning. *arXiv* **2021**, arXiv:2107.00233.
24. Zhou, T.; Konukoglu, E. FedFA: Federated Feature Augmentation. *arXiv* **2023**, arXiv:2301.12995.
25. Li, W.; Milletarì, F.; Xu, D.; Rieke, N.; Hancox, J.; Zhu, W.; Baust, M.; Cheng, Y.; Ourselin, S.; Cardoso, M.J.; et al. Privacy-preserving federated brain tumour segmentation. In Proceedings of the International Workshop on Machine Learning in Medical Imaging-MICCAI 2019, Shenzhen, China, 13–17 October 2019; Springer International Publishing: Cham, Switzerland, 2019; pp. 133–141. [CrossRef]
26. Lo, J.; Yu, T.T.; Ma, D.; Zang, P.; Owen, J.P.; Zhang, Q.; Wang, R.K.; Beg, M.F.; Lee, A.Y.; Jia, Y.; et al. Federated learning for microvasculature segmentation and diabetic retinopathy classification of OCT data. *Ophthalmol. Sci.* **2021**, *1*, 100069. [CrossRef]
27. Vaid, A.; Jaladanki, S.K.; Xu, J.; Teng, S.; Kumar, A.; Lee, S.; Somani, S.; Paranjpe, I.; De Freitas, J.K.; Wanyan, T.; et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach. *JMIR Med. Inform.* **2021**, *9*, e24207. [CrossRef]
28. Luc, P.; Couprie, C.; Chintala, S.; Verbeek, J. Semantic segmentation using adversarial networks. *arXiv* **2016**, arXiv:1611.08408.
29. Xue, Y.; Xu, T.; Zhang, H.; Long, L.R.; Huang, X. SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation. *Neuroinformatics* **2018**, *16*, 383–392. [CrossRef] [PubMed]
30. Lei, B.; Xia, Z.; Jiang, F.; Jiang, X.; Ge, Z.; Xu, Y.; Qin, J.; Chen, S.; Wang, T.; Wang, S. Skin lesion segmentation via generative adversarial networks with dual discriminators. *Med. Image. Anal.* **2020**, *64*, 101716. [CrossRef] [PubMed]

31. Nguyen, D.C.; Ding, M.; Pathirana, P.N.; Seneviratne, A.; Zomaya, A.Y. Federated learning for COVID-19 detection with generative adversarial networks in edge cloud computing. *IEEE Internet Things J.* **2021**, *9*, 10257–10271. [CrossRef]
32. Rasouli, M.; Sun, T.; Rajagopal, B. Fedgan: Federated generative adversarial networks for distributed data. *arXiv* **2020**, arXiv:2006.07228.
33. Fan, C.; Liu, P. Federated generative adversarial learning. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV 2020), Nanjing, China, 16–18 October 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 3–15. [CrossRef]
34. Zhang, Y.; Qu, H.; Chang, Q.; Liu, H.; Metaxas, D.; Chen, C. Training federated gans with theoretical guarantees: A universal aggregation approach. *arXiv* **2021**, arXiv:2102.04655.
35. Yang, D.; Xu, Z.; Li, W.; Myronenko, A.; Roth, H.R.; Harmon, S.; Xu, S.; Turkbey, B.; Turkbey, E.; Wang, X.; et al. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. *Med. Image. Anal.* **2021**, *70*, 101992. [CrossRef]
36. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. In Proceedings of the Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5769–5779. [CrossRef]
37. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
38. Jun, M.; Cheng, G. COVID-19 CT Lung and Infection Segmentation Dataset | Kaggle. 2020. Available online: https://www.kaggle.com/andrewmvd/covid19-ct-scans/ (accessed on 9 June 2022).
39. Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation. *Med. Phys.* **2021**, *48*, 1197–1210. [CrossRef]
40. Ma, J.; Wang, Y.; An, X.; Ge, C.; Yu, Z.; Chen, J.; Zhu, Q.; Dong, G.; He, J.; He, Z.; et al. MosMedData: Chest CT Scans with COVID-19 Related Findings Dataset. *Med. Phys.* **2020**, *48*, 1197–1210. [CrossRef]
41. Jenssen, H.B. COVID-19 Radiology-Data Collection and Preparation for Artificial Intelligence. 2020. Available online: http://medicalsegmentation.com/covid19/ (accessed on 2 June 2022).
42. COVID-19 DATABASE | SIRM. Available online: https://sirm.org/category/senza-categoria/covid-19/ (accessed on 2 June 2022).
43. Liu, Q.; Dou, Q.; Heng, P.A. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference 2020, Lima, Peru, 4–8 October 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 475–485. [CrossRef]
44. Litjens, G.; Toth, R.; Ven, W.; Hoeks, C.; Kerkstra, S.; Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J.; et al. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Med. Image. Anal.* **2014**, *18*, 359–373. [CrossRef]
45. Lemaître, G.; Martí, R.; Freixenet, J.; Vilanova, J.C.; Walker, P.M.; Meriaudeau, F. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. *Comput. Biol. Med.* **2015**, *60*, 8–31. [CrossRef]
46. NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures. Available online: https://wiki.cancerimagingarchive.net/display/Public/NCI-ISBI+2013+Challenge+-+Automated+Segmentation+of+Prostate+Structures/ (accessed on 20 March 2023).
47. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Proceedings of the Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 1195–1204. [CrossRef]
48. Miyato, T.; Maeda, S.I.; Koyama, M.; Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1979–1993. [CrossRef]
49. Verma, V.; Kawaguchi, K.; Lamb, A.; Kannala, J.; Solin, A.; Bengio, Y.; Lopez-Paz, D. Interpolation consistency training for semi-supervised learning. *Neural Netw.* **2022**, *145*, 90–106. [CrossRef]
50. Zhang, Y.; Yang, L.; Chen, J.; Fredericksen, M.; Hughes, D.P.; Chen, D. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention-MICCAI 2017, Quebec City, QC, Canada, 11–13 September 2017; Springer International Publishing: Cham, Switzerland, 2017; pp. 408–416. [CrossRef]
51. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.