Federated Synthetic Learning from Multi-institutional and Heterogeneous Medical Data

Qi Chang

Rutgers University <https://orcid.org/0000-0003-2146-785X>

Zhennan Yan

SenseBrain Research

Hui Qu

Adobe

Han Zhang

Google Brain

Lohendran Baskaran

Department of Cardiovascular Medicine, National Heart Centre, Singapore

Shaoting Zhang

SenseTime Research <https://orcid.org/0000-0002-0057-2944>

Dimitris Metaxas (✉ dnm@cs.rutgers.edu)

Rutgers University

Article

Keywords: deep learning, synthetic database, machine learning, data analytics

Posted Date: November 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-965641/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Federated Synthetic Learning from Multi-institutional and Heterogeneous Medical Data

Qi Chang^{1,6}, Zhennan Yan^{2,6}, Hui Qu^{1*}, Han Zhang^{1†}, Lohendran Baskaran³, Shaoting Zhang⁴, and Dimitris N. Metaxas¹

¹*Department of Computer Science, Rutgers University, Piscataway, New Jersey, USA*

²*SenseBrain Research, Princeton, NJ, USA*

³*Department of Cardiovascular Medicine, National Heart Centre Singapore, and Duke-National University Of Singapore*

⁴*SenseTime Research and Shanghai AI Laboratory, Shanghai, China*

⁵*These authors contributed equally: Qi Chang and Zhennan Yan*

Statistically and information-wise adequate data plays a critical role in training a robust deep learning model. However, collecting sufficient medical data to train a centralized model is still challenging due to various constraints such as privacy regulations and security. In this work, we develop a novel privacy-preserving federated-discriminator GAN, named FedD-GAN, that can learn and synthesize high-quality and various medical images regardless of their type, from heterogeneous datasets residing in multiple data centers whose data cannot be transferred or shared. We trained and evaluated FedD-GAN on three essential classes of medical data, each involving different types of medical images: cardiac CTA, brain MRI, and histopathology. We show that the synthesized images using our method have better quality than using a standard federated learning method and are realistic and accurate enough to train accurate segmentation models in downstream tasks. The segmentation model trained on the synthetic data only is comparable to that trained on an all-in-one real-image dataset shared from multiple data centers if possible. FedD-GAN can learn to generate a scalable and diverse synthetic database without compromising data privacy. This synthetic database could help to boost machine learning techniques in medical data analytics.

*Present affiliation: Adobe, CA, USA

†Present affiliation: Google Brain, CA, USA

Introduction

Advances in machine learning (ML), especially deep learning (DL) ¹, have shown promising results in automatic medical data analytics in recent years, which could benefit medical diagnosis and treatment. For example, various DL techniques have been adopted in many medical image analysis tasks with superior performances to conventional ML methods ^{2,3}. It is well-known that statistically and information-wise sufficient data samples play a critical role in training a successful ML model ^{4,5}. Deep learning, which involves more complex models of millions or billions of parameters, usually requires much more data samples than classical ML algorithms to overcome the overfitting issue. For example, the large-scale image dataset, ImageNet ⁶ with 14 million images, is one of the keys to the success of modern deep learning methods used in computer vision research. Since heterogeneity often appears in healthcare data due to varied populations, pathologies, environments, procedures, and treatment protocols, a large-scale medical dataset that can cover the data diversity is necessary for training a robust and accurate DL model. But in the medical image domain, the largest publicly available dataset DeepLesion ⁷ is only about 0.2% of the size of the ImageNet. Although it contains over 32 thousand axial slices, it is only about 0.01% of the annually acquired CT scans in the U.S. ⁸.

Currently, medical datasets for large-scale data analytics is commonly achieved by voluntary data sharing from multiple institutions and even multiple countries ⁹. However, collecting these multi-institutional or even multi-national medical data is very challenging. The multi-institutional collaborations in the medical domain are often confronted with the requirement of increasing data diversity and data size in the presence of protocol prohibiting data sharing to a centralized location. The privacy regulations, such as HIPAA ^{10,11}, EU GDPR ^{12,13}, and the Institutional Review Board (IRB) ¹⁴ approval process, restrict the distribution and acquisition of medical data to protect the patients' private information. For instance, the patient data are not allowed to leave their country in America, European Union, and many other countries ^{15,16}. Furthermore, common practices in centralized data sharing need to anonymize and transmit data to a site where data can be analyzed by training some ML models. However, data anonymization is not perfect for protecting privacy ¹⁷ and needs to be carefully and specifically implemented due to different medical data formats and standards.

Federated learning (FL) ^{18,19} provides a potential solution to this challenge. It learns a model across multiple decentralized devices or servers holding local data samples without sharing or copying them to a server for processing. This technique has attracted more and more attention in healthcare ²⁰⁻²³. For example, recent studies demonstrated successful application of FL in segmenting whole tumor from T2-FLAIR images²⁴, classifying breast density in mammography ²⁵, and detecting COVID-19 in CT images ²⁶. The main limitation of the classical FL is that it is inefficient to adapt to architecture updates because it usually learns a globally shared model for a specific task. The rapid evolution of deep learning techniques in recent years very often result in recently well-trained models become outdated and under-perform in a very short time as new techniques, such as new network modules, loss functions, or optimizers, are invented. Thus, it will take lots of effort from all participating centers to keep training new models in the presence of data

sharing restrictions. Due to different regulations and data usage agreements, the currently available data may not be accessible in the future. Therefore, we may not be able to take advantage of the new techniques when some dataset is offline. Instead of directly training a task-specific model, e.g. segmentation or classification model, a recent study tried to use a generative adversarial network (GAN) ²⁷ to learn joint data distribution in federated setting (FLGAN) ²⁸. This learned generative model could synthesize similar data samples to the real ones for downstream uses. However, such federated-learned GAN may not be effective when the data in different centers are highly heterogeneous ^{29,30}. Because learning GAN is not a trivial task ³¹ and simply averaging the parameters or gradients to train a shared GAN could be more unstable if each local GAN sees different data distribution. In real-world scenarios, besides the population difference, the medical imaging protocols in different centers are not identical due to various data acquisition protocols and related technical issues in practice. For example, multi-modality imaging is commonly used in tumor detection and disease diagnosis. However, selecting the appropriate protocol depends on the indication for the exam and the patient history ^{32,33}. It is challenging for participating data centers to collect homogeneous multi-modality data since some modalities may be missing. Some existing methods, like cross-modality domain adaption ³⁴, image translation ³⁵, and missing modality completion ³⁶ could solve a similar problem, but they require centralized training data. Federated transfer learning ^{37,38} aims to handle this challenging problem in which data parties share only a partial overlap in the user or feature space, but it has not been extensively explored.

To cope with the above limitations, we propose a Federated-Discriminator Generative Adversarial Network (FedD-GAN), as shown in Fig. 1. The FedD-GAN consists of one Generator located in a central server and several Discriminators that are privately hosted by different entities. The communications only happen between the generator and multiple distributed discriminators. As a result, the learned generator can not only learn the joint distribution of heterogeneous data, but also generate a scalable synthetic database from the learned distribution for public research purposes, which can benefit the development and evolution of new techniques. We apply our method to the learning of three categories of distributed heterogeneous data. The well-trained centralized generator can capture the joint distribution of distributed data, and produce high-quality synthetic images. We utilized the synthetic images to train segmentation models and test on real cases, which achieved comparable performance to models trained with real data. This implies that our method can be used as a data provider that benefits a lot of downstream tasks.

Results

Summary of datasets and experimental settings We collected three categories of heterogeneous datasets for cardiac computed tomography angiography (CTA), brain magnetic resonance imaging (MRI), and histopathological images. The characteristics of these datasets are summarized in table 1, which shows large heterogeneity, e.g., various subject numbers, voxel spacings, scanners, organs, hospitals/centers, and so on.

In the first category, we collected three public cardiac CTA datasets acquired from globally different institutes: the Whole Heart Segmentation (WHS) challenge dataset³⁹, Automated Segmentation of Coronary Arteries (ASOCA) challenge 2020 dataset⁴⁰, and MICCAI Coronary Artery Tracking (CAT08) Challenge 2008 dataset⁴¹. We included all 60 CTA images in WHS dataset with a mean voxel size of $0.44 \times 0.44 \times 0.6 \text{ mm}^3$, all 32 images in CAT08 with a mean voxel size of $0.32 \times 0.32 \times 0.4 \text{ mm}^3$, all 40 ASOCA training samples with a mean voxel size of $0.4 \times 0.4 \times 0.625 \text{ mm}^3$. Because these data were acquired for diagnosing or analyzing cardiovascular diseases, we used the multi-structure heart mask images as the input (condition) to the generator to synthesize corresponding CTA images. The substructures include (1) the left ventricle blood cavity (LV), (2) the right ventricle blood cavity (RV), (3) the left atrium blood cavity (LA), (4) the right atrium blood cavity (RA), (5) the myocardium of the left ventricle (Myo), (6) the ascending aorta (aorta), (7) the pulmonary artery (PA)⁴².

In the second category, we used 210 studies of glioblastoma (GBM/HGG) from the Brain Tumor Segmentation (BraTS) challenge^{43,44}, which were acquired with different clinical protocols and various scanners from 19 different institutions including the Center for Biomedical Image Computing and Analytics (CBICA), the Cancer Imaging Archive (TCIA), and other contributors (Other). We used 170 for training and validation and 40 for testing. Each case comprises four MRI modalities, including native (T1), T1 with gadolinium enhancing contrast (T1c), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). The ground truth annotation contains three types of tumor sub-regions including tumor core, enhancing tumor, and edema. We used the tumor sub-region labels and brain skull contour as the input to the generator.

At last, we chose the multi-organ nuclei image dataset (**Nuclei**)⁴⁵ as the third category. This dataset contains 30 digital microscopic tissue images from 30 patients and around 17,000 annotated nuclear boundaries in total (including both epithelial and stromal nuclei). These images of size 1000×1000 came from 18 different hospitals spanning seven organs. The training set has 16 images from the breast, kidney, liver, and prostate (4 images per organ). The testing set has 14 images from 7 organs (2 images per organ), and three organs are not in the training set: bladder, colon, and stomach. In the training of FedD-GAN, we used the nuclear boundary annotations as the input to the generator. We partitioned the training data into 4 sites based on the organ types.

We conducted three different learning tasks for the three categories of data. After the learning of FedD-GAN, the generator can generate images and form a synthetic database. We assessed the synthetic image quality quantitatively by computing the DistFID score (see definition in Meth-

ods). We can select the best generator model with the lowest DistFID score. Then, we used the synthetic images in a downstream task, i.e., training an image segmentation model, to implicitly evaluate the image quality. We assumed that if the quality of synthetic data is good enough, then the segmentation model would perform similarly to the model learned from the real data. For the downstream segmentation tasks, we withheld 20% of the training subjects as a validation subset to select the best segmentation models. The evaluation metrics for the cardiac CTA and brain tumor MRI tasks were subject-level volumetric DICE, 95% Hausdorff Distance (HD95), and average Surface Distance (SD). The nuclei segmentation were validated by object-level DICE⁴⁶ and Aggregated Jaccard Index (AJI)⁴⁵. See Methods for the details of the preprocessing steps and the metrics.

Learning from multi-center cardiac CTA data Firstly, we learned FedD-GAN on multi-center cardiac CTA images. Taken a multi-structural heart label image as the input, the goal of the FedD-GAN is to learn from the joint distribution of data samples at isolated private data centers by adversarial learning strategy to obtain a central generator that can generate realistic CTA images. In this task, we simulated three data centers: WHS, CAT08, and ASOCA.

Figure 2 shows some examples of real vs synthetic images. The synthetic images by FLGAN²⁸, a standard federated synchronized and averaged GAN model, had an DistFID score of 72.05, while our FedD-GAN can generate better-quality images with an DistFID of 61.09. Table 2 and Figure 6 report the quantitative performances of different methods in the downstream segmentation task. Figure 9 compares their segmentation results for several samples. In the cardiac sub-structure segmentation, Real-All represents the rough upper bound of the segmentation performance because it utilizes all real samples shared from all data centers. Such centralized learning is ideal for a good segmentation model but may encounter many challenges due to privacy regularizations. The models trained from the local real data, including Real-WHS, Real-CAT08, and Real-ASOCA, underperform the Real-All model by a large margin ranging from 7.5% to 25%. It is because the data variance in a local dataset is limited.

By using the synthetic images generated by FedD-GAN, the segmentation model can obtain comparable performance to the Real-All model and statistically better than using any local real dataset in all three metrics ($p < 0.05$). It is also significantly better ($p = 0.0001$) than FLGAN. It indicates that our learning strategy works better than the parameter averaging algorithm for the heterogeneous data across different centers. If directly training a segmentation model from real data in the federated learning fashion (Real-FedSeg)¹, the performance was similar to the ideal Real-All results. However, our method can be more flexible and friendly for future updates, while the FedSeg model is fixed for a specific task once trained.

Learning from multi-center brain MRI data Secondly, we validated on the multi-modality MRI datasets by a small adjustment in the network architecture (detailed in Methods). Taken tumor sub-region labels and the brain skull as the input, FedD-GAN learns to generate realistic multi-modal brain MRI images. The central generator produces a 4-channel output corresponding to the four

¹<https://fedsegment.github.io/home/>

MRI modalities. For this task, the training data was partitioned into three sites: CBICA, TCIA, and Other.

Figure 3 shows some samples of the real vs synthetic images. The synthetic images by FLGAN had an DistFID score of 42.09, while FedD-GAN can synthesize images with an DistFID score of 38.77. Table 3 summarizes the quantitative results of the downstream segmentation task, and Figure 7 shows the comparison in detail. Figure 10 visualizes some examples of segmentation results.

We found that by only using synthesized images by FedD-GAN, the segmentation model revealed a statistically better result than the model trained on Real-CBICA or Real-Other data centers in terms of DICE, HD, and SD ($p < 0.05$). The model trained on real data of TCIA performed similarly to the models by FedSeg and FedD-GAN. Most likely, it is because the TCIA set is relatively large (about 50% size of Real-All) so that the data distribution may cover more variance than its peers. Similar to the cardiac CTA task, the segmentation model learned from synthetic images of FLGAN significantly underperformed the models trained from the real local datasets (Real-CBICA and Real-TCIA) and FedD-GAN. For this task, FedD-GAN obtained similar performance as the FedSeg method ($p > 0.05$) while the latter does not have the flexibility for future upgrades.

Furthermore, the synthetic database of FedD-GAN could be used as data augmentation for any private data center to boost its segmentation performance. Specifically, by comparing the fourth and the second sections in Table 3, we noticed that after introducing synthetic images into any data center to train segmentation, the DICE scores were significantly improved ($p < 0.05$), and the distance metrics were also improved ($p < 0.05$) for center CBICA.

Next, we evaluated the robustness of the methods in a more heterogeneous setting where the MRI modalities are misaligned across data centers. To simplify the problem, we removed one modality from each data center. As a result, the CBICA center missed all Flair images, the TCIA center missed the T1c modality, and the Other center missed the T2 images. It was a challenging scenario as the data features were not the same across data centers, which may require some complex federated transfer learning methods to resolve. We adjusted the FedSeg and FedD-GAN methods for learning from this missing-modality task and reported the segmentation results in Table 4 as Hetero-FedSeg and Hetero-FedD-GAN. In Hetero-FedSeg, we set the pixel values of missing modalities to 0 and trained the FedSeg. In Hetero-FedD-GAN, we simply adjusted the discriminators in the architecture (see Methods for details). As shown in Table 4 and Figure 4, our method can handle this challenging problem with very small performance loss by “completing” the missing modalities with synthesized images, while a trivially adjusted FedSeg failed to achieve satisfying segmentation from missing-modality data.

Learning from multi-organ histopathology data At last, we evaluated FedD-GAN on multi-organ histopathology images. In this nuclei task, we simulated four data centers, each containing data of a single organ: breast, liver, kidney, and prostate. During training, the original images were cropped to smaller tiles (see Methods for details). Figure 5 shows examples of real vs syn-

thetic images. The proposed method generated much better synthetic images with an DistFID score of 159.60 than the FLGAN with an DistFID score of 234.16. Table 5 compares the average performance of the downstream segmentation task by different methods, and Figure 8 plots the distribution of the results. Figure 11 shows some segmentation samples. Our FedD-GAN was on par with the Real-All model and FedSeg. Compared with those models using the local data (Real-breast, Real-liver, Real-kidney, and Real-prostate), our method achieved significantly better results in at least one of the metrics.

The FLGAN method is also reported in this table and is statistically worse than FedD-GAN in terms of both DICE and AJI ($p < 0.01$). The results imply that our method could generate more realistic and diverse synthetic image samples than FLGAN, which can benefit the downstream ML task.

Discussion

Learning from multi-institutional and heterogeneous data is a challenging but practical problem in large-scale medical image analysis. In this study, we developed a GAN-based federated-learning architecture, FedD-GAN. It can learn to synthesize medical images from heterogeneous/non-independent-and-identically-distributed (non-IID) datasets. It consists of a central generator and multiple distributed discriminators. We demonstrated that the proposed framework could learn from various data across data centers, with varied data sizes and images acquired from different scanners, subjects, and organs. FedD-GAN can be used to build a centralized and scalable synthetic database for downstream machine learning tasks without access to any private information.

The proposed learning framework is general and could be used in various clinically useful applications. We have implemented FedD-GAN in a distributed-training environment in the FedML library⁴⁷ and showed its effectiveness in heterogeneous/non-IID data settings in various and practical use cases, including cardiac CTA images, multi-modality brain MR images, and histopathological images. The synthetic images can be assessed quantitatively and selected automatically by using a novel DistFID metric (described in Methods). From the qualitative visualizations and quantitative DistFID score, we can see that the synthetic images generated by the learned generator have higher image quality than the FLGAN²⁸. In addition, we validated the synthetic images in segmentation tasks by using them as training data and testing on real cases to compare with models trained on real data. The results of the downstream segmentation tasks showed that the segmentation models learned from only synthetic images can achieve close performance to the models trained on the all-in-one real dataset collected by copying from multiple institutions.

Our method was inspired by the federated learning, particularly the horizontal FL^{19,38}, to resolve the privacy and data collection concerns in medical image analysis. In a horizontal FL setting, it is assumed that data in different centers is partitioned by examples, all of which have the same features. During the training, the standard FL generally learns a global shared model by repeatedly deploying the same model to each center and aggregating the model parameters or gradients from all data centers in each round. There are several critical differences between our method and the horizontal FL in the architecture design and the training process. Firstly, our method does not learn the task-specific model directly from collaborative learning. Instead, our goal is to learn a synthetic data generator from the data distributions at all participating centers. Therefore, once the generator is trained, it can be used to build a sizeable synthetic database for multiple specific downstream tasks without the burden of federated learning of different models numerous times. Secondly, our method deploys different models at the central and distributed local centers. Specifically, there is only a generator on the central server, and each data center only has its own unshared local discriminator. On the contrary, the standard federated learning assumes an identical network structure in each party, for example, the FLGAN²⁸ has both shared generator and discriminator in all agents, the FedAvg-GAN⁴⁸ has shared discriminator in all agents. Thirdly, our method does not need to exchange model parameters between server and data centers as standard FL or between the peer centers as in MD-GAN⁴⁹. Thus, it avoids the additional communication cost and concerns of privacy and security between data centers.

The architecture of FedD-GAN is basically a conditional GAN^{50,51} with multiple discriminators. GANs²⁷ are still being intensively developed. For example, StyleGANs^{52,53} were proposed to improve image quality in unconditional image modeling and enable intuitive, scale-specific control of the synthesis. Some other studies explored different architectures involving multiple generators^{54,55} or multiple discriminators^{56,57} to overcome the mode collapsing problem. Another recent work adopted multiple generators and multiple discriminators for medical image fusion⁵⁸. In this study, the purpose of distributed multiple discriminators is to learn from multi-institutional and heterogeneous data sets. The exchanged information between the central generator and each local discriminator includes the annotations, generated images, and back-propagated feedback. The annotations can be collected once from data sites to the central server before the training process begins. The learning process is similar to the FedSGD scheme⁵⁹. The central generator sends the fake images to each local discriminator and collects the discriminators' feedback to update its parameters. Since FedD-GAN deploys different models in each party, it is not trivial to adopt some delay-aggregation strategies, such as FedAVG⁵⁹. Although FedAVG can reduce communication cost, it may not perform well when learning in heterogeneous environments^{29,30}, which is also demonstrated in our experiments by comparing the results of the FLGAN (implemented with FedAVG) and ours.

The missing-modality experiment for the brain MRI task in Results can be considered a federated transfer learning³⁸ scenario because only partial feature space (some of the MR modalities) is shared between data centers. Federated transfer learning is a challenging yet practical concept, however, only a few studies^{37,60} focus on it so far. Our framework can still learn a robust generator in this challenging situation, which does not require sophisticated transfer learning techniques^{61,62}.

This work contains completely new results and a substantial extension of methodology, analysis, and conclusions over our previous work⁶³. In our preliminary study⁶³, we validated the method in IID data settings for single-modality images, and the method was implemented in a simulated stand-alone environment. After training, it required manual selection of the generator model by visualizing its generated samples in different epochs in order to build the synthetic database. The 2D images were simply assumed independent of each other in the segmentation tasks and the results were measured in 2D metrics.

In the implementation of FedD-GAN, noise is only provided in the form of dropout on several layers of the generator network in both training and inference. When building the synthetic database, we observed only minor stochasticity in the synthetic images given the same input. This observation is similar in *pix2pix*⁵¹. Therefore, by default we just generated the same amount of synthetic images as the total number of the real data in all participating centers. Since in many medical applications, the size of the annotated dataset is often a concern, we adopted two ways to make FedD-GAN generate scalable synthetic database with diverse images. One is applying random transformations to the input, such as scaling, shift, flip, and rotation, to generate more varied images (see examples in Supplementary Fig. 1, Fig. 2, and Fig. 3). The other one is utilizing multiple generator models saved at different training epochs with the smallest DistFID

scores. We found that in the histopathology task, when the synthetic database becomes twice the size, the performance of the down-stream task was improved (Dice: from 0.789 to 0.805, AJI: from 0.528 to 0.552). This indicates implicitly the good quality and diversity of the synthetic images. However, further increasing the synthetic database of histopathology images to more than twice the size can not benefit the downstream models. Increasing synthetic database for the cardiac CTA and brain MRI tasks cannot improve downstream task either. A possible reason is that as the synthetic database becomes larger some repeating patterns and artifacts in synthetic images may also accumulate and cause overfitting problems in the training of downstream models. When the overall size of real data is small (like in histopathology task), scaling up the synthetic database could bring more benefit than harm until the useful information in the synthetic database 'saturate'. Thus, we conducted a manual evaluation by asking a radiologist to distinguish between 200 randomly picked pairs of real images and synthetic images for cardiac CTA. Each pair was shuffled randomly. It turns out that the radiologist can effectively tell fake from real images (over 90% accuracy) by looking for some unnatural artifacts and clues in background (unconditioned) regions. For example, the synthetic images may have artificial "grid" artifacts around the top and bottom of the heart (where the annotated area is very small) or unrealistic clues in areas of the front ribs and back spine (see the last column of Figure 2). To bridge the gap between fake and real images, recent studies have investigated on improving realism^{52,53,64} and reducing mode collapse^{54,55,57,65,66} of synthetic images in the learning of GANs. By incorporating these new techniques into our method, the quality and diversity of the synthetic images may be further upgraded leading to an infinite annotated dataset generator. Such a labeled data factory⁶⁷ could benefit many downstream tasks in the medical domain by reducing expensive human efforts.

In terms of privacy and security, a recent work showed that the gradients can be used to recover the original pixel-wise image data (deep leakage) in a classical multi-node training, where the same model is shared⁶⁸. Although such attack could work on FedSeg and FLGAN, which learn a shared model, it may not work on FedD-GAN. Because FedD-GAN does not have any model shared in any party, the central server and all participants learn different models and no parameter is exchanged. We notice that the synthetic images tend to have different contexts from the original images especially in the unconditioned regions (see Figures 2, 3, 4, 5 and more examples in Supplementary). These random contexts can help to ensure privacy by not being able to produce identical samples to the real ones. It is a limitation of FedD-GAN, since it only takes sparse annotations as the input and is not able to control the overall contexts. A better future technique may generate realistic but different image by explicitly controlling some high-level attributes⁵². FedD-GAN could also be enhanced by incorporating additional security mechanisms like secure aggregation⁶⁹ and differential privacy⁷⁰ for real-life applications. Note that in the experiments, FLGAN and FedSeg do not have additional security implementation either for a fair comparison of performance.

FedD-GAN could be further improved by optimizing the communication cost while preserving performance and privacy. Regarding the communication cost, the amount of data exchanges in our method is proportional to the size of overall training data. In contrast, in FedSeg and FLGAN that based on FedAVG, the exchange size relies on the number of the learnable parameters of the desired

network. Therefore, for a large-scale dataset, our approach may need heavier communication than the FedAVG. However, the proposed method should be more favorable in terms of scalability, image quality, and downstream task performance, especially for non-IID data. It is because the federated learning in the medical domain is more like a cross-silo setting¹⁹ with fewer clients and more powerful computing capacities and reliable connections than in a cross-device setting¹⁹.

In future work, we would like to investigate more complex practical problems with more heterogeneous situations, for example, learning from more pathological conditions, learning from both labeled and unlabeled data, etc. It is also of interest to combine imaging data with other forms of electronic health record, e.g., lab results and radiology reports, into one learning framework. Furthermore, other forms of controlling factors for the generator (instead of segmentation masks) and different downstream tasks could be investigated, for example, using bounding boxes or global labels to generate data for detection or classification, or even text-to-image generation⁷¹. Another research direction would be lifelong learning⁷², in which a model can be continuously learned for different tasks leading to a more general artificial intelligence. These problems are practical and yet more challenging.

Conclusion The contributions are in four aspects. Firstly, the proposed framework can be an excellent solution to the privacy issue in collaborative studies that involve multiple institutes because the generator can learn the joint data distribution of numerous datasets in different entities without direct accessing or storing patients' private image data. Secondly, the well-trained generator can be used as an image provider to create a publicly accessible database of synthetic images without sensitive information. Since the generator learned from heterogeneous data to produce images with varied appearances, this centralized synthetic database can be ready to benefit researchers. In addition, since the private-sensitive data may not always be accessible, our generator may become the only source when the data centers are offline. Thirdly, our method provides a solution for the adaptation to future architecture updates. In order to take advantage of quickly evolved techniques, such as loss functions⁷³, network architectures^{74–77} and so on, task-specific federated learning, e.g. the FedSeg, needs a complete retrain on all private datasets. However, since our method can produce a synthetic database, exploration or refinement of the task-specific models can be done locally without worry about the loss of access to the distributed real datasets. Last but not least, our framework is easily extensible. We have demonstrated its application for both single- and multi-modality image generation in this study. It can also handle the missing-modality problem by providing a feasible way to learn from incomplete multi-modality datasets in the real scenario and forming a hybrid central data provider for the downstream tasks. As another example, the generator could be learned in a life-long scheme to generate different images across use cases⁷².

Methods

Data collection and processing We evaluated our method on three different medical image databases, including multi-center cardiac CTA, brain MRI, and multi-organ histopathology, to demonstrate the generalization of the proposed method.

For the **Cardiac CTA** data, we collected three public CTA datasets acquired from globally different institutes: WHS dataset^{39,78}, ASOCA challenge 2020 dataset⁴⁰, MICCAI CAT08 Challenge dataset⁴¹. The WHS data have manually annotated labels of seven whole heart substructures. We generated the annotations of the same substructures for CAT08 and ASOCA datasets by using a state-of-the-art whole heart segmentation algorithm⁷⁹ in the SenseCare research platform⁸⁰ and manually correcting gross errors. All the cardiac CTA data were resampled to isotropic 0.8 mm resolution. We used 200 and 1000 as the window level and width to transfer the Hounsfield units to intensity values in our experiments.

For the brain tumor MR images, we used the Brain Tumor Segmentation challenge (**BraTS18**) 2018 dataset^{43,44,81}. We selected the 210 multimodal magnetic resonance imaging (MRI) scans of glioblastoma (GBM/HGG) from the challenge training set to conduct our experiments. All modalities have been aligned to a common space and resampled to 1mm isotropic resolution⁴⁴.

For the histopathology images, we used the multi-organ nuclear segmentation dataset (**Nuclei**)⁴⁵ containing 30 images of 30 patients and about 17,000 nuclear boundary annotations (including both epithelial and stromal nuclei). These images of size 1000×1000 came from 18 different hospitals spanning seven organs. In the preprocessing step⁸², we first performed color normalization⁸³ for all images. Then, each image was divided into 16 (4×4) overlapping tiles. In the training of the FedD-GAN, we employed a tile size of 286×286 . In the training of the segmentation model, we used a tile size of 256×256 .

Network architecture Our proposed FedD-GAN is comprised of only one central generator and multiple distributed discriminators located in different local nodes. An overview of the proposed architecture is shown in Figure 1. The central generator, denoted as G , takes task-specific inputs (e.g., segmentation masks in our use case) and generates synthetic images to fool the discriminators. Let N denote the number of participating entities that collaborate in the learning framework, and $S_j = \{(x_i^j, y_i^j) : i = 1, \dots, s^j\}$ denote the local private dataset at the j -th entity, where x is an auxiliary variable representing annotation, such as a class label or segmentation mask, and y is the corresponding real image data. The local discriminators, denoted as $D_j, j \in \{1, \dots, N\}$, learn to differentiate between the local real images y_i^j and the synthetic images $\hat{y}_i^j = G(x_i^j)$ generated from G based on x_i^j . Our architecture ensures that D_j deployed in the j -th medical entity only has access to its local dataset while not sharing any real image data outside the entity. Only synthetic images, annotations, and losses are transferred between the central generator and the distributed discriminators during the learning process.

Central generator: For segmentation tasks, the central generator is designed to generate im-

ages based on input masks so that the synthetic image and corresponding mask can be used as a pair to train a segmentation model. Here, an encoder-decoder resnet⁵¹, is adopted for G . It consists of nine residual blocks⁸⁴, two stride-2 convolutions for downsampling, and two transposed convolutions for upsampling. All non-residual convolutional layers are followed by batch normalization⁸⁵ and the ReLU activation. All convolutional layers use 3×3 kernels except the first and last layers that use 7×7 kernels.

Distributed discriminators: In our framework, each discriminator has the same structure as that in PatchGAN⁵¹. The discriminator classifies each of the overlapping patches of the input image as real or fake. Such architecture assumes patch-wise independence of pixels in a Markov random field fashion^{51,86}, and the patch is large enough (70×70) to capture the difference in geometrical structures such as background and tumors.

The generator can learn the joint distribution of multiple isolated datasets through adversarial learning. Then, it can be used as an image provider to generate training samples for some downstream tasks. Assuming the distribution of synthetic images, $p_{\hat{y}}$, is the same or similar to that of the real images, p_{data} , we can generate one large unified dataset which approximately equals to the union of all the datasets in medical entities. In this way, all private image data from each entity are utilized without sharing. To evaluate the synthetic images, we use the generated samples in segmentation tasks to illustrate the effectiveness of the proposed FedD-GAN.

Objective function The FedD-GAN is based on the conditional GAN⁵⁰. The objective function is:

$$\begin{aligned} & \min_G \max_{D_1:D_N} V(D_{1:N}, G) \\ &= \sum_{j \in [N]} \pi_j \{ \mathbb{E}_{x \sim s_j(x)} [\mathbb{E}_{y \sim p_{data}(y|x)} \log D_j(y|x) \\ & \quad + \mathbb{E}_{\hat{y} \sim p_{\hat{y}}(\hat{y}|x)} \log(1 - D_j(\hat{y}|x))] \} \end{aligned} \quad (1)$$

The goal of D_j is to maximize Eq. (1), while G minimizes it. In this way, the learned $G(x)$ with maximized $D(G(x))$ can approximate the real data distribution $p_{data}(y|x)$ and D cannot tell ‘fake’ data from real. x follows a distribution $s(x)$. In this paper, We assume that the joint distribution $s(x) = \sum_{j \in [N]} \pi_j s_j(x)$, where $s_j(x)$ is marginal distribution of j -th dataset and π_j is the prior distribution. In the experiment, we set $s_j(x)$ be a uniform distribution and $\pi_j \propto |S_j|$, resulting in a uniform distribution $s(x)$. For each sub-distribution, there is a corresponding discriminator D_j which only receives data generated from prior $s_j(x)$. Similar to previous works^{51,87}, we incorporate noises by using Dropout⁸⁸ at several layers of the generator G in both training and inference, instead of providing a Gaussian noise as input to the generator.

The losses of D_j and G are defined in Eq. (2) and Eq. (3), respectively.

$$L_{D_j} = \frac{1}{m} \sum_{i=1}^m [-\log D_j(y_i^j | x_i) - \log(1 - D_j(\hat{y}_i^j | x_i))] , \quad (2)$$

445

$$L_G = \frac{1}{Nm \sum \pi_j} \sum_{j=1}^N \pi_j \sum_{i=1}^m [\log(1 - D_j(\hat{y}_i^j | x_i)) + \lambda_1 L_1(y_i^j, \hat{y}_i^j) + \lambda_2 L_P(y_i^j, \hat{y}_i^j)]. \quad (3)$$

446 where m is the minibatch size. The L_G contains perceptual loss (L_P)⁸⁹ and L_1 loss besides the
 447 adversarial loss. In this study, G and D_j are not on the same server and thus the equation (3) needs
 448 to be split into two parts (Eq. (4) and Eq. (5)) in order to back-propagate the loss to G .

$$L_{G_j} = \frac{1}{m} \sum_{i=1}^m [\log(1 - D_j(\hat{y}_i^j | x_i)) + \lambda_1 L_1(y_i^j, \hat{y}_i^j) + \lambda_2 L_P(y_i^j, \hat{y}_i^j)]. \quad (4)$$

449

$$\nabla_{\hat{y}} = \frac{1}{N \sum \pi_j} \sum_{j=1}^N \pi_j [\nabla_{\hat{y}^j}], \quad (5)$$

450 where $\nabla_{\hat{y}^j} = \frac{\partial L_{G_j}}{\partial \hat{y}^j}$ is computed at node D_j based on loss in Eq. (4) and then sent back to G
 451 for aggregation (Eq. (5)). The learning process is summarized in Algorithm 1. We trained 200
 452 epochs for all tasks and updated each discriminator once in each training iteration. The gradient-
 453 based updates can adopt different gradient-based learning rules. We used Adam optimizer⁹⁰ with
 454 learning rate 0.0002 in our experiments.

455 **Extension for multi-modality data** For a use case of multi-modality data, assuming c modalities,
 456 the local data center j has a set of multi-modality image $\mathbf{y}_i^j = (y_{i,1}^j, \dots, y_{i,c}^j)$ associated with each
 457 label image x_i^j . A simple way of handling the multi-modality image in our framework would be
 458 treating the c modalities of one sample as a c -channel image. Thus the only change needed is the
 459 number of channels of the input layer of D and an output layer of G . In this setting, the learning
 460 task of D could be easier and converge very fast since different modalities have different contrast
 461 patterns, and more information can be used to differentiate the real and the 'fake' data. However,
 462 the task of G may become more challenging to learn. It is because, on one hand, the G needs to
 463 learn more complex data distribution to generate multiple modalities with different contrasts. On
 464 the other hand, the easily-learned D may learn some trivial discriminative features and thus cannot
 465 provide helpful feedback to G to guide its learning.

466 To balance the task difficulty of the G and D 's, we extend our framework by deploying multiple
 467 discriminators at each entity. Every single modality has its discriminator in one data center, and the
 468 G receives losses from the multiple D s for a multi-modality data sample. In this way, each D can
 469 focus on learning discriminative features for one specific modality and provide more meaningful
 470 feedback to G . The objective function can be extended from Eq. 1 as:

$$\begin{aligned} & \min_G \max_{D_{1:N}^{1:c}} V(D_{1:N}^{1:c}, G) \\ &= \sum_{j \in [N]} \pi_j \{ \mathbb{E}_{x \sim s_j(x)} \sum_{k=1}^c [\mathbb{E}_{y_k \sim p_{data}(y_k|x)} \log D_{j,k}(y_k|x) \\ &+ \mathbb{E}_{\hat{y}_k \sim p_{\hat{y}}(\hat{y}_k|x)} \log(1 - D_{j,k}(\hat{y}_k|x))] \}, \end{aligned} \quad (6)$$

where $D_{j,k}$ represents the discriminator for the k -th modality at the center j .

Besides, another advantage of the proposed multi-modality framework is that it enables learning from missing modality data. Let C_j denote the set of index of available modality for center j , if data center j misses the s -th modality for example, then $C_j = \{1, \dots, s-1, s+1, \dots, c\}$. In this case, center j only needs to deploy $c-1$ discriminators during the learning. The learning process has no difference except that it only collects losses of available discriminators for C_j to update the G and only use a subset of the synthetic images $\{\hat{y}_k^j | k \in C_j\}$ to update the corresponding $\{D_{j,k} | k \in C_j\}$ in center j . Because the discriminators for different modalities in different entities are all independent, the G can still learn to generate all modalities, assuming that the missing modality in one center is available in some other data centers. The loss function of D is the same, while the loss function of G can be adjusted as the following:

$$L_G = \frac{1}{Nm \sum \pi_j} \sum_{j=1}^N \pi_j \sum_{i=1}^m \sum_{k \in C_j} [\log(1 - D_{j,k}(\hat{y}_{i,k}^j | x_i)) + \lambda_1 L_1(y_{i,k}^j, \hat{y}_{i,k}^j) + \lambda_2 L_P(y_{i,k}^j, \hat{y}_{i,k}^j)]. \quad (7)$$

After training, the learned G can act as a synthetic image provider to generate multi-modality images from the conditional variable, a mask image. As a result, it can also be used for missing modality completion. For instance, if a data center has data $(y_1, \dots, y_{s-1}, y_{s+1}, y_c)$ with the s -th modality missing and the corresponding mask image x , we can use the synthetic image $\hat{y}_s = G_s(x)$ as a substitute. Our approach is different from the existing methods that predict the target modality from another modality^{35,36} in the sense that it can generate multiple modalities to handle randomly missing modality problems, and thus does not require a specific model for specific modality pair for the input and output.

Distributed FID for image quality measurement The Frechet Inception Distance (FID)⁹¹ has been widely used to evaluate the image quality by calculating the distance between the statistics of feature vectors of the real and generated images. The definition of FID is:

$$FID = \|\mu_1 - \mu_2\|^2 + Tr(\sigma_1 + \sigma_2 - 2 * sqrt(\sigma_1 * \sigma_2)), \quad (8)$$

where μ_1 and μ_2 refer to the feature-wise mean of the real and generated images, σ_1 and σ_2 are the covariance matrices for the real and generated feature vectors, Tr refers to the trace operation in linear algebra.

Though FID is an ideal metric to find the best model when train a GAN^{19,53}, we are unable to compute one FID score in federated learning because a joint set of the isolated real data does not exist. Therefore, we propose a new metric named distributed FID (DistFID) to calculate the weighted average distance between each real dataset and the synthetic database. The DistFID is defined as:

$$DistFID = \sum_i^N w_i (\|\mu_1^i - \mu_2\|^2 + Tr(\sigma_1^i + \sigma_2 - 2 * sqrt(\sigma_1^i * \sigma_2))) \quad (9)$$

in which each of the N entities host a dataset S_i of size $|S_i|$ with feature statistics (μ_1^i, σ_1^i) . The weight of each center $w_i = \frac{|S_i|}{\sum_{i=1}^N |S_i|}$. We validated the consistency between the FID and DistFID scores in Supplementary Figure 4.

Learning of downstream task In this study, we used segmentation as the downstream machine learning task. After obtaining a well-learned image generator from the FedD-GAN, we can generate synthetic medical images from mask (label) images. In our experiments, to fairly compare the effect of the synthetic images with the real samples, we adopted the same U-Net⁹² as the segmentation model to learn on different sets of 2D images. During training the segmentation, we withheld 20% samples from the training data as the validation set to select the model with the best DICE score to test. We used Adam optimizer with a learning rate of 0.01 to learn segmentation in our experiments. For the cardiac CTA and brain MRI segmentation tasks, a combination of cross-entropy (CE) and DICE was used as the loss function. For the nuclear segmentation task, CE loss was used. We inferred every 2D image in testing and for the cardiac CTA and brain MRI data we computed the 3D metrics by stacking up the 2D images for the same subject, which are reported in Results.

Quantitative metrics The Dice score (Dice) and 95% quantile of Hausdorff distance (HD95) are adopted to evaluate the segmentation performance on cardiac CTA and BraTS18⁴⁴. The Dice score measures the overlap between ground-truth mask G and segmented result S . It is defined as

$$Dice(G, S) = \frac{2|G \cap S|}{|G| + |S|} \quad (10)$$

The Hausdorff Distance (HD) evaluates the distance between boundaries of ground-truth and segmented masks:

$$HD(G, S) = \max\left\{\sup_{x \in \partial G} d(x, \partial S), \sup_{y \in \partial S} d(y, \partial G)\right\} \quad (11)$$

where ∂ means the boundary operation, $d(x, \partial S) = \inf_{y \in \partial S} \|x - y\|_2$ is minimum distance from point x to surface ∂S and $\sup$ represents the supremum and $\inf$ the infimum. Because the Hausdorff distance is sensitive to outliers in G or S , we use the 95% quantile Hausdorff distance (HD95):

$$HD95(G, S) = \max\left\{\sup_{x \in \partial G}^{95} d(x, \partial S), \sup_{y \in \partial S}^{95} d(y, \partial G)\right\}, \quad (12)$$

where the $\sup^{95}$ is the 95% maximum value. In addition, we report the average Surface Distance (SD) as follows:

$$SD(G, S) = \frac{1}{2} \left\{ \frac{1}{|\partial G|} \sum_{x \in \partial G} d(x, \partial S) + \frac{1}{|\partial S|} \sum_{y \in \partial S} d(y, \partial G) \right\}. \quad (13)$$

For nuclei segmentation, we utilize the object-level Dice⁴⁶ and the Aggregated Jaccard Index (AJI)⁴⁵:

$$AJI = \frac{\sum_{i=1}^{n_G} |G_i \cap S(G_i)|}{\sum_{i=1}^{n_G} |G_i \cup S(G_i)| + \sum_{k \in K} |S_k|} \quad (14)$$

529 where $S(G_i)$ is the segmented object that has maximum overlap with G_i with regard to Jaccard
530 index, K is the set containing segmentation objects that have not been assigned to any ground-truth
531 object.

532 **Data Availability**

533 The cardiac datasets include publicly available MM-WHS (<https://zmiclab.github.io/projects/mmwhs/>),
534 ASOCA (<https://asoca.grand-challenge.org/>), and CAT08 (<http://coronary.bigr.nl/centerlines/>). The
535 WHS masks for ASOCA and CAT08 data can be downloaded at <https://rutgers.app.box.com/folder/147492331099>.
536 The brain data are available through BraTS 2018 Challenge (<https://www.med.upenn.edu/sbia/brats2018.html>).
537 The nuclei data⁴⁵ are now part of the MoNuSeg training set (<https://monuseg.grand-challenge.org/Data/>).

538 **Code Availability**

539 FedD-GAN is implemented in Python 3.7 using PyTorch framework 1.6.0⁹³. It is based on FedML
540 library⁴⁷. The source code of FedD-GAN and segmentation used in this study can be found at:
541 <https://rutgers.app.box.com/folder/147492331099>.

542 **Acknowledgements**

543 We thank Qing Xia and Wenji Wang for helping to generate segmentation masks for CAT08
544 and ASOCA datasets using their method⁷⁹. This work was supported in part by grants from
545 National Science Foundation (IIS-1703883, CNS-1747778, CCF-1733843, IIS-1849238-825536,
546 ARO MURI-Z8424104-440149).

Figure 1: The overview of the FedD-GAN architecture and workflow. The architecture contains one central generator and multiple distributed discriminators, each located in a medical entity. The generator takes a conditional input (segmentation masks in our experiments) and outputs synthetic data. Each discriminator learns to differentiate between its own real images and synthetic images received from the generator, and then sends back the gradients. The generator updates by the adversarial learning. At last, the well-trained generator can be used as an image provider to build a synthetic database for downstream machine learning tasks, e.g., segmentation in this study.

Table 1: Summary of datasets.

Cardiac CTA	WHS ³⁹		ASOCA ⁴⁰		CAT08 ⁴¹		
Train Subjects (Image#)	20(3031)		32(4642)		26(3568)		
Test Subjects (Image#)	40(6360)		8(1180)		6(756)		
Avg Spacing (<i>mm</i> ³)	0.44×0.44×0.6		0.4×0.4×0.625		0.32×0.32×0.4		
Scanner	Philips		Unknown		Siemens		
BraTS18 ^{43,44}	CBICA		TCIA		Other		
Train Subjects(Image#)	69(4638)		85(5736)		14(975)		
Test Subjects(Image#)	19(393)		17(1165)		6(1173)		
Modalities	T1, T1c, T2 and Flair						
Nuclei ⁴⁵	Breast	Liver	Kidney	Prostate	Bladder	Colon	Stomach
Train Subjects(Nuclei#)	4(1508)	4(1906)	4(4866)	4(1634)	0	0	0
Test Subjects(Nuclei#)	2(707)	2(838)	2(716)	2(766)	2(743)	2(726)	2(2556)

Figure 2: Five examples of real and synthetic cardiac images. (a) The mask images of cardiac sub-structures are the input for image synthesise. (b) Corresponding real CTA images. (c) The synthetic images by FLGAN’s generator. (d) The synthetic images by FedD-GAN’s generator.

Figure 3: Two examples of real and synthetic multi-modality brain images. (a) The label of the brain tumor. Columns (b)-(e): T1, T2, FLair and T1c images. Rows 1 and 4 show the real images. Rows 2 and 5 show the synthetic images generated by FLGAN’s generator. Rows 3 and 6 show the synthetic images generated by FedD-GAN’s generator.

Figure 4: FedD-GAN can learn from missing modality data. Here shows one sample from each center. (a) The label of brain tumor. Columns (b)-(e): T1, T2, FLair and T1c images. Row 1, 3 and 5 show the real images with one missing modality during the training of FedD-GAN. Row 2, 4 and 6 show the synthetic images of FedD-GAN after training. The red dash boxes indicate the missing images (top) and the completed images by FedD-GAN (bottom).

Figure 5: Four examples of real and synthetic nuclei images for different organs in each row. (a) The label images with nuclear boundaries. (b) Corresponding real images. (c) Synthetic images by FLGAN from input (a). (d) Synthetic images by FedD-GAN.

Table 2: Quantitative results in the cardiac CTA segmentation in terms of the average score of seven sub-structures. In the first column, 'Real-' indicates the model was trained from original real images, otherwise the model was trained from synthetic images. Real-All merges together all data of WHS, CAT08, and ASOCA sites. Real-FedSeg directly learns a segmentation model from real data in distributed centers.

Data/Method	Average Dice $\uparrow$	Average HD95(mm) $\downarrow$	Average SD(mm) $\downarrow$
Real-All	0.906 $\pm$ 0.037	6.89 $\pm$ 5.69	1.67 $\pm$ 0.97
Real-FedSeg	0.913 $\pm$ 0.042	10.57 $\pm$ 8.26	1.38 $\pm$ 0.82
Real-WHS	0.831 $\pm$ 0.086	13.81 $\pm$ 9.34	3.37 $\pm$ 2.15
Real-CAT08	0.675 $\pm$ 0.145	33.24 $\pm$ 12.12	9.28 $\pm$ 4.80
Real-ASOCA	0.838 $\pm$ 0.093	18.86 $\pm$ 14.47	3.94 $\pm$ 2.87
FLGAN	0.709 $\pm$ 0.153	37.33 $\pm$ 15.47	6.14 $\pm$ 2.89
FedD-GAN	0.864 $\pm$ 0.068	13.23 $\pm$ 7.93	2.85 $\pm$ 1.66

Table 3: Quantitative results of whole tumor segmentation in multi-modal brain MRI. In the first column, 'Real-' indicates the model was trained from original real images, otherwise the model was trained from synthetic images. Real-All merges together all data of CBICA, TCIA, and Other sites. Real-FedSeg directly learns a segmentation model from real data in distributed centers. 'Syn + Real-' represents synthetic data augmentation by adding the synthetic data from FedD-GAN.

Data/Method	Dice $\uparrow$	HD95(mm) $\downarrow$	SD(mm) $\downarrow$
Real-All	0.862 $\pm$ 0.128	7.56 $\pm$ 10.90	1.48 $\pm$ 2.02
Real-FedSeg	0.839 $\pm$ 0.157	9.24 $\pm$ 12.42	1.75 $\pm$ 1.69
Real-CBICA	0.801 $\pm$ 0.142	27.11 $\pm$ 25.12	4.54 $\pm$ 4.42
Real-TCIA	0.823 $\pm$ 0.117	10.32 $\pm$ 10.53	1.90 $\pm$ 1.19
Real-Other	0.765 $\pm$ 0.167	15.61 $\pm$ 13.39	2.81 $\pm$ 2.36
FLGAN	0.736 $\pm$ 0.197	19.23 $\pm$ 17.19	3.50 $\pm$ 2.69
FedD-GAN	0.829 $\pm$ 0.128	11.50 $\pm$ 12.82	1.99 $\pm$ 1.86
Syn + Real-CBICA	0.841 $\pm$ 0.156	9.56 $\pm$ 13.37	1.91 $\pm$ 3.07
Syn + Real-TCIA	0.854 $\pm$ 0.093	10.50 $\pm$ 12.29	1.71 $\pm$ 1.30
Syn + Real-Other	0.824 $\pm$ 0.126	17.83 $\pm$ 24.08	2.81 $\pm$ 3.46

Table 4: Quantitative comparison of segmentation models learned by FedSeg and from synthetic data of FedD-GAN. 'Real-FedSeg' indicates the segmentation model was trained from real data in distributed centers. 'FedD-GAN' indicates the model was trained from the synthetic images generated by FedD-GAN. 'Hetero-' represents the adjusted methods learning from missing-modality data.

Data/Method	Dice $\uparrow$	HD95(mm) $\downarrow$	SD(mm) $\downarrow$
Real-FedSeg	0.839 $\pm$ 0.157	9.24 $\pm$ 12.42	1.75 $\pm$ 1.69
Hetero-FedSeg	0.353 $\pm$ 0.144	65.84 $\pm$ 10.36	16.26 $\pm$ 3.66
FedD-GAN	0.829 $\pm$ 0.128	11.50 $\pm$ 12.82	1.99 $\pm$ 1.86
Hetero-FedD-GAN	0.7956 $\pm$ 0.193	18.37 $\pm$ 19.44	4.00 $\pm$ 9.61

Table 5: Quantitative results in the nuclear segmentation. In the first column, 'Real-' indicates the model was trained from original real images, otherwise the model was trained from synthetic images. Real-All merges together all data of local sites. Real-FedSeg directly learns a segmentation model from real data in distributed centers.

Data/Method	Dice $\uparrow$	AJI $\uparrow$
Real-All	0.801 $\pm$ 0.053	0.549 $\pm$ 0.091
Real-FedSeg	0.807 $\pm$ 0.052	0.537 $\pm$ 0.094
Real-breast	0.765 $\pm$ 0.074	0.442 $\pm$ 0.112
Real-liver	0.757 $\pm$ 0.071	0.364 $\pm$ 0.135
Real-kidney	0.721 $\pm$ 0.093	0.442 $\pm$ 0.128
Real-prostate	0.770 $\pm$ 0.060	0.508 $\pm$ 0.077
FLGAN	0.707 $\pm$ 0.072	0.370 $\pm$ 0.128
FedD-GAN	0.805 $\pm$ 0.049	0.552 $\pm$ 0.075

Figure 6: The detailed box plots for cardiac CTA segmentation. Different models are trained using either real data (All, WHS, CAT08, ASOCA, FedSeg) or synthetic data (FLGAN and FedD-GAN). The first two rows are for DICE metric. Eight figures show the results of average score, and scores of LV, RV, LA, RA, Myo, Aorta, and PA structures, respectively. The row 3 and 4 are HD95 results. The last two rows are SD results.

Figure 7: The detailed box plots for whole tumor segmentation in multi-modal brain MRI. (a) The comparison of different methods using real data, synthetic data or combined data. (b) The comparison between the FedSeg using real data and the centralized model by using the synthetic database of FedD-GAN. By using all 4 modality data our FedD-GAN had comparable results as the Real-FedSeg. By using missing-modality data, Hetero-FedD-GAN achieved much more robust results than Hetero-FedSeg.

Figure 8: The detailed box plots for nuclear segmentation in terms of accuracy, DICE, and AJI metrics. The testing set contains the same four organs as the training set and three unseen organs. Learning from the synthetic images of FedD-GAN can achieve comparable segmentation performance as directly learning from all-in-one data collecting from all centers.

Figure 9: Visualization of three results in cardiac CTA segmentation. (a) and (b) show the original CTA image and corresponding ground truth. The other columns show results of different models trained from real data or synthetic data.

Figure 10: Visualization of three results in brain tumor segmentation. (a) and (b) show the original MR image and corresponding ground truth. The other columns show results of different models trained from real data or synthetic data.

Figure 11: Visualization of two results in nuclear segmentation. (a) and (b) show the original histopathology image and corresponding ground truth. The other columns show results of different models trained from real data or synthetic data. Due to limited space, results of Real-liver, Real-kidney and Real-prostate are not shown.

Algorithm 1: Algorithm of learning FedD-GAN

Input: the number of discriminators (clients) N , initial generator and discriminator parameters θ_G, θ_{D_j} ($j \in \{1, \dots, N\}$), minibatch size m , number of iterations R_{max} , sample distribution $s_j(x)$ and size of dataset $|S_j|$ at node j

```

1 for  $j = 1 : N$  do
2    $D_j$  sends its available auxiliary variables  $\{x_i^j | i = 1, \dots, |S_j|\}$  to generator  $G$ ;
3 end
4 for  $r = 1 : R_{max}$  do
5   for  $j = 1 : N$  do
6      $G$  samples minibatch of  $m$  variables  $x_i^j \sim s_j(x)$  as input to generate  $m$  fake data
        $\hat{y}_i^j = G(x_i^j)$ , and sends them to node  $D_j$ ;
7      $D_j$  finds corresponding samples  $y_i^j$  and updates  $\theta_{D_j}$  according to the loss Eq. 2;
8     fix  $\theta_{D_j}$  to compute gradients for  $\hat{y}^j$  based on Eq. 4 and send to  $G$ ;
9   end
10  for  $j = 1 : N$  do
11     $G$  receives gradients  $\nabla_{\hat{y}^j}$  from  $D_j$ ;
12  end
13   $G$  aggregates gradients (Eq. 5) with  $\pi_j = |S_j|$  and updates  $\theta_G$ ;
14 end
15 return The learned models  $\theta_G$  and  $\theta_{D_j}$  ( $j \in \{1, \dots, N\}$ )

```

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).
2. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Medical Image Analysis* **42**, 60–88 (2017).
3. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annual review of biomedical engineering* **19**, 221–248 (2017).
4. Domingos, P. M. A few useful things to know about machine learning. *Communications of the ACM* **55**, 78–87 (2012).
5. Najafabadi, M. M. *et al.* Deep learning applications and challenges in big data analytics. *Journal of big data* **2**, 1–21 (2015).
6. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (Ieee, 2009).
7. Yan, K., Wang, X., Lu, L. & Summers, R. M. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging* **5**, 036501 (2018).
8. Papanicolas, I., Woskie, L. R. & Jha, A. K. Health care spending in the united states and other high-income countries. *Jama* **319**, 1024–1039 (2018).
9. Hasell, J. *et al.* A cross-country database of covid-19 testing. *Scientific data* **7**, 1–7 (2020).
10. for Disease Control, C., Prevention *et al.* Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services. *MMWR: Morbidity and mortality weekly report* **52**, 1–17 (2003).
11. Annas, G. J. *et al.* Hipaa regulations-a new era of medical-record privacy? *New England Journal of Medicine* **348**, 1486–1490 (2003).
12. Regulation, P. General data protection regulation. *Intouch* (2018).
13. Goddard, M. The eu general data protection regulation (gdpr): European regulation that has a global impact. *International Journal of Market Research* **59**, 703–705 (2017).
14. Bankert, E. A. & Amdur, R. J. *Institutional review board: Management and function* (Jones & Bartlett Learning, 2006).
15. Seddon, J. J. & Currie, W. L. Cloud computing and trans-border health data: Unpacking us and eu healthcare regulation and compliance. *Health Policy and Technology* **2**, 229–241 (2013).
16. Kerikmäe, M. S. T. Challenges in the e-health regulatory policy. *HOW DEEP IS YOUR LAW? BREXIT. TECHNOLOGIES. MODERN CONFLICTS* **367** (2017).
17. Schwarz, C. G. *et al.* Identification of anonymous mri research participants with face-recognition software. *New England Journal of Medicine* **381**, 1684–1686 (2019).

18. McMahan, H. B., Moore, E., Ramage, D. & y Arcas, B. A. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629* (2016).
19. Kairouz, P. *et al.* Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977* (2019).
20. Brisimi, T. S. *et al.* Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics* **112**, 59–67 (2018).
21. Rieke, N. *et al.* The future of digital health with federated learning. *NPJ digital medicine* **3**, 1–7 (2020).
22. Xu, J. *et al.* Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research* **5**, 1–19 (2021).
23. Kaissis, G. *et al.* End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence* **3**, 473–484 (2021).
24. Sheller, M. J. *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports* **10**, 1–12 (2020).
25. Roth, H. R. *et al.* Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 181–191 (Springer, 2020).
26. Dou, Q. *et al.* Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study. *NPJ digital medicine* **4**, 1–11 (2021).
27. Goodfellow, I. *et al.* Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680 (2014).
28. Rasouli, M., Sun, T. & Rajagopal, R. Fedgan: Federated generative adversarial networks for distributed data. *arXiv preprint arXiv:2006.07228* (2020).
29. Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine* **37**, 50–60 (2020).
30. Li, T. *et al.* Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127* (2018).
31. Salimans, T. *et al.* Improved techniques for training gans. *Advances in neural information processing systems* **29**, 2234–2242 (2016).
32. Brown, A. D. & Marotta, T. R. Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. *Journal of the American Medical Informatics Association* **25**, 568–571 (2018).

33. Ellingson, B. M. *et al.* Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro-oncology* **17**, 1188–1198 (2015).
34. Chen, C., Dou, Q., Chen, H., Qin, J. & Heng, P.-A. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 865–872 (2019).
35. Zhang, Z., Yang, L. & Zheng, Y. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition* (2018).
36. Cai, L., Wang, Z., Gao, H., Shen, D. & Ji, S. Deep adversarial learning for multi-modality missing data completion. In *24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1158–1166 (2018).
37. Gao, D. *et al.* Privacy-preserving heterogeneous federated transfer learning. In *2019 IEEE International Conference on Big Data (Big Data)*, 2552–2559 (IEEE, 2019).
38. Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* **10**, 1–19 (2019).
39. Zhuang, X. & Shen, J. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical Image Analysis* **31**, 77–87 (2016).
40. Gharleghi, R., Samarasinghe, G., Sowmya, A. & Beier, S. Automated segmentation of coronary arteries challenge. <https://asoca.grand-challenge.org> (2020).
41. Schaap, M. *et al.* Standardized evaluation methodology and reference database for evaluating coronary artery centerline extraction algorithms. *Medical Image Analysis* **13**, 701–714 (2009).
42. Zhuang, X. *et al.* Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical Image Analysis* **58**, 101537 (2019).
43. Bakas, S. *et al.* Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data* **4**, 170117 (2017).
44. Bakas, S. *et al.* Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629* (2018).
45. Kumar, N. *et al.* A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging* **36**, 1550–1560 (2017).
46. Qu, H. *et al.* Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Transactions on Medical Imaging* **39**, 3655–3666 (2020).
47. He, C. *et al.* Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518* (2020).

48. Augenstein, S. *et al.* Generative models for effective ml on private, decentralized datasets. *arXiv preprint arXiv:1911.06679* (2019).
49. Hardy, C., Le Merrer, E. & Sericola, B. MD-GAN: Multi-discriminator generative adversarial networks for distributed datasets. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 866–877 (IEEE, 2019).
50. Mirza, M. & Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
51. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134 (2017).
52. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410 (2019).
53. Karras, T. *et al.* Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8110–8119 (2020).
54. Ghosh, A., Kulharia, V., Namboodiri, V. P., Torr, P. H. & Dokania, P. K. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8513–8521 (2018).
55. Hoang, Q., Nguyen, T. D., Le, T. & Phung, D. Mgan: Training generative adversarial nets with multiple generators. In *International conference on learning representations* (2018).
56. Nguyen, T. D., Le, T., Vu, H. & Phung, D. Dual discriminator generative adversarial nets. *arXiv preprint arXiv:1709.03831* (2017).
57. Mordido, G., Yang, H. & Meinel, C. microbatchgan: Stimulating diversity with multi-adversarial discrimination. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3061–3070 (2020).
58. Huang, J. *et al.* Mgmdbgcn: Medical image fusion using multi-generator multi-discriminator conditional generative adversarial network. *IEEE Access* **8**, 55145–55157 (2020).
59. McMahan, B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282 (PMLR, 2017).
60. Chen, Y., Qin, X., Wang, J., Yu, C. & Gao, W. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems* **35**, 83–93 (2020).
61. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**, 1345–1359 (2009).

62. Zhuang, F. *et al.* A comprehensive survey on transfer learning. *Proceedings of the IEEE* **109**, 43–76 (2020).
63. Chang, Q. *et al.* Synthetic learning: Learn from distributed asynchronized discriminator GAN without sharing medical image data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 13856–13866 (2020).
64. Esser, P., Rombach, R. & Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883 (2021).
65. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).
66. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).
67. Zhang, Y. *et al.* Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10145–10155 (2021).
68. Zhu, L. & Han, S. Deep leakage from gradients. In *Federated learning*, 17–31 (Springer, 2020).
69. Damgård, I., Pastro, V., Smart, N. & Zakarias, S. Multiparty computation from somewhat homomorphic encryption. In *Annual Cryptology Conference*, 643–662 (Springer, 2012).
70. Dwork, C., Roth, A. *et al.* The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* **9**, 211–407 (2014).
71. Zhang, H., Koh, J. Y., Baldridge, J., Lee, H. & Yang, Y. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 833–842 (2021).
72. Qu, H. *et al.* Learn distributed gan with temporary discriminators. In *European Conference on Computer Vision*, 175–192 (Springer, 2020).
73. Ma, J. *et al.* Loss odyssey in medical image segmentation. *Medical Image Analysis* **71**, 102035 (2021).
74. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision (3DV)*, 565–571 (IEEE, 2016).
75. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818 (2018).

- 715 76. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks.
716 In *International Conference on Machine Learning*, 6105–6114 (PMLR, 2019).
- 717 77. Khan, S. *et al.* Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169* (2021).
- 718 78. Zhuang, X. Challenges and methodologies of fully automatic whole heart segmentation: a
719 review. *Journal of healthcare engineering* **4**, 371–407 (2013).
- 720 79. Wang, W. *et al.* Few-shot learning by a cascaded framework with shape-constrained pseudo la-
721 bel assessment for whole heart segmentation. *IEEE Transactions on Medical Imaging* (2021).
- 722 80. Duan, Q. *et al.* Sensecare: A research platform for medical image informatics and interactive
723 3d visualization. *arXiv preprint arXiv:2004.07031* (2020).
- 724 81. Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (BRATS).
725 *IEEE Transactions on Medical Imaging* **34**, 1993–2024 (2014).
- 726 82. Qu, H., Yan, Z., Riedlinger, G. M., De, S. & Metaxas, D. N. Improving nuclei/gland in-
727 stance segmentation in histopathology images by full resolution neural network and spatial
728 constrained loss. In *International Conference on Medical Image Computing and Computer-*
729 *Assisted Intervention*, 378–386 (Springer, 2019).
- 730 83. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE*
731 *Computer Graphics and Applications* **21**, 34–41 (2001).
- 732 84. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE*
733 *Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
- 734 85. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing
735 internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
- 736 86. Li, C. & Wand, M. Precomputed real-time texture synthesis with markovian generative adver-
737 sarial networks. In *European Conference on Computer Vision*, 702–716 (Springer, 2016).
- 738 87. Mathieu, M., Couprie, C. & LeCun, Y. Deep multi-scale video prediction beyond mean square
739 error. *arXiv preprint arXiv:1511.05440* (2015).
- 740 88. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A sim-
741 ple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*
742 **15**, 1929–1958 (2014).
- 743 89. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for real-time style transfer and super-
744 resolution. In *European Conference on Computer Vision* (2016).
- 745 90. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint*
746 *arXiv:1412.6980* (2014).

- 747 91. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Gans trained by a two
748 time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information*
749 *Processing Systems* **30** (2017).
- 750 92. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical im-
751 age segmentation. In *International Conference on Medical image computing and computer-*
752 *assisted intervention*, 234–241 (Springer, 2015).
- 753 93. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Ad-*
754 *vances in neural information processing systems* **32**, 8026–8037 (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FedDGANNaturesupv1.0Oct12.pdf](#)