Variation-Aware Federated Learning With Multi-Source Decentralized Medical Image Data

Zengqiang Yan, Jeffry Wicaksana, Zhiwei Wang, Xin Yang, and Kwang-Ting Cheng, *Fellow, IEEE*

***Abstract*—Privacy concerns make it infeasible to construct a large medical image dataset by fusing small ones from different sources/institutions. Therefore, federated learning (FL) becomes a promising technique to learn from multi-source decentralized data with privacy preservation. However, the cross-client variation problem in medical image data would be the bottleneck in practice. In this paper, we propose a variation-aware federated learning (VAFL) framework, where the variations among clients are minimized by transforming the images of all clients onto a common image space. We first select one client with the lowest data complexity to define the target image space and synthesize a collection of images through a privacy-preserving generative adversarial network, called PPWGAN-GP. Then, a subset of those synthesized images, which effectively capture the characteristics of the raw images and are sufficiently distinct from any raw image, is automatically selected for sharing with other clients. For each client, a modified CycleGAN is applied to translate its raw images to the target image space defined by the shared synthesized images. In this way, the cross-client variation problem is addressed with privacy preservation. We apply the framework for automated classification of clinically significant prostate cancer and evaluate it using multi-source decentralized apparent diffusion coefficient (ADC) image data. Experimental results demonstrate that the proposed VAFL framework stably outperforms the current horizontal FL framework. As VAFL is independent of deep learning architectures for classification, we believe that the proposed framework is widely applicable to other medical image classification tasks.**

***Index Terms*—Cross-client variation, federated learning, medical image analysis, prostate cancer classification.**

Manuscript received July 11, 2020; revised October 21, 2020; accepted November 18, 2020. Date of publication November 24, 2020; date of current version July 20, 2021. This work was supported by Hong Kong General Research Fund (GRF) 16203319, a research grant from WeBank and NSFC 61872417. *(Corresponding author: Xin Yang.)*

Zengqiang Yan and Kwang-Ting Cheng are with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, Hong Kong (e-mail: zengqiangyan@ust.hk; timcheng@ust.hk).

Jeffry Wicaksana and Zhiwei Wang are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, Hong Kong (e-mail: jwicaksana@connect.ust.hk; wangzw_pk@126.com).

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

Digital Object Identifier 10.1109/JBHI.2020.3040015

I. Introduction

FEDERATED learning (FL) [1], which can learn from multi-source decentralized data without data sharing or collection, has drawn great attention recently. Based on the distribution characteristics of data, current FL frameworks can be roughly classified into the following three categories [2].

Horizontal Federated Learning: Horizontal FL is the most widely used FL framework and is designed for the scenario that datasets of clients share the same feature space but differ in samples. Shokri *et al.* [3] proposed a collaboratively deep learning scheme by allowing participants to train independently and share partial updates of parameters. Smith *et al.* [4] proposed to allow multiple sites to train for separate tasks while sharing knowledge and preserving security. McMahan *et al.* [1] built a client-server structure which allows models built at different clients to collaborate at the server to build a global federated model.

Vertical Federated Learning: Vertical FL is designed for the case where datasets share the same samples but differ in feature space. In [5], [6], a vertical FL framework was proposed to train a privacy-preserving logistic regression model. Feng *et al.* [7] proposed a multi-participant multi-class vertical federated learning (MMVFL) framework, by transferring a multi-view learning approach into the vertical FL setting.

Federated Transfer Learning: Federated transfer learning is for the case where datasets differ not only in samples but also in feature space. It can be regarded as transfer learning under the constraints of FL. Peng *et al.* [8] implemented unsupervised domain adaptation under the FL setting for natural image data. The variations between the source clients and the target client in the feature space were minimized through adversarial learning. Similarly, Ju *et al.* [9] proposed a federated transfer learning framework for EEG signal classification, where the variations among clients in the feature space were reduced by using maximum mean discrepancy (MMD).

When dealing with multi-source decentralized medical image data, as clients usually would share the same task but different cases/patients, definitely horizontal FL is the most suitable framework. A typical horizontal FL framework is shown in Fig. 1. Supposed there are N participating clients, at time t the server sends the global model parameters w_G^t to each client. Each client i performs local model update $w_i^{t+1} \leftarrow clientupdate(w_G^t)$ solely based on its local data D_i. All clients send their updates to the server and the parameters of the global

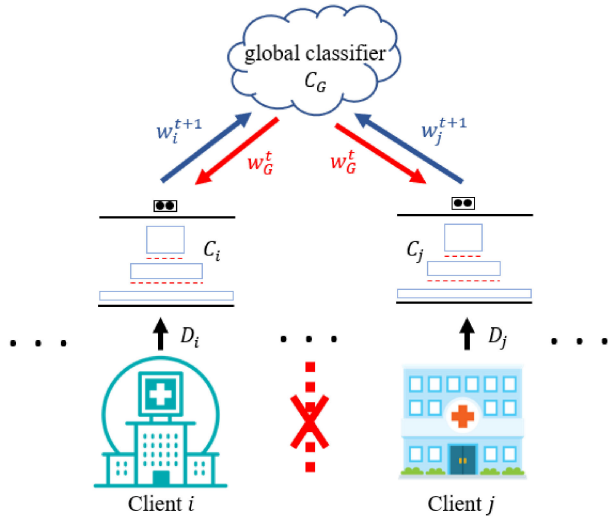

Fig. 1. A typical horizontal FL framework.

model at time $t+1$ are updated accordingly via an aggregation function $w_G^{t+1} = \sum_{i=1}^{N} p_i w_i^{t+1}$, where p_i is the weight of client i. The training process iterates until convergence. As only updated parameters rather than the raw training data are shared, FL can learn from multi-source decentralized data with privacy preservation.

Despite the success of FL in privacy preservation, it encounters serious limitations while dealing with multi-source decentralized medical image data, due to the following reasons:

1) Limited annotated data: Different from natural image data where usually a large amount of annotated data can be used for training such as ImageNet [10] ($\sim 1,500,000$ images), CIFAR-10 ($\sim 60,000$ images), etc., annotated medical image data can be very limited such as In-Breast [11] (410 images from 115 patients), DRIVE [12] (40 images), etc., due to the significant efforts required in data acquisition and annotation. The lack of training data can lead to poor transferability, which explains why performance degradation is quite common in the cross-dataset evaluation as reported in [13]–[15]. As a result, in FL, the parameters updated at each client can be highly sensitive to the characteristics of its local training data.
2) Limited scale: Theoretically, the number of clients in FL can be very large. As the cost of acquisition and annotation of natural image data is much lower than that of medical image data, millions of clients can be expected when applying FL with natural image data; however, the corresponding number of medical clients in a community is usually much lower (in the order of 10 s).
3) Variations among clients: Though data acquisition usually follows similar procedures, the same type of medical image data acquired by different institutions/hospitals could exhibit different characteristics. For example, there exist significant variations between the fundus image datasets DRIVE [12] and STARE [16] for retinal vessel segmentation, between the ultrasound image datasets BUSI [17] and Dataset B [18] for breast tumor segmentation, etc. Although the variation problem among medical image data may not be as severe as that of natural image data, it could be more serious considering that the amount of annotated medical image data is much more limited. Consequently, the parameters updated by clients can vary significantly, leading to bad convergence after aggregation.

Based on the above observations, the cross-client variation problem, compounded with the limitation of available training data of each client and the limited number of participating clients, becomes the bottleneck of FL in dealing with multi-source decentralized medical image data. To address the cross-client variation problem, we propose a variation-aware federated learning (VAFL) framework. The key idea is to translate the raw training images of all clients to a predefined image space under the FL setting. The first step is to define the target image space by evaluating the data complexity of all clients, and the training images of the client with the lowest complexity are chosen to define the target image space. To share the defined image space with other clients, we synthesize a collection of images through a privacy-preserving generative adversarial network (*i.e.*, PPWGAN-GP) based on the chosen client's raw images and automatically select a subset of the synthesized images which can effectively capture the characteristics of the raw images and are sufficiently distinct from any raw image. According to the privacy designs of PPWGAN-GP, sharing those selected synthesized images with other clients is safe without leaking any privacy of the raw images. Then, for each client, its raw training images are translated to the target image space defined by the shared synthesized images. In this way, as the raw images of all clients are translated to the same image space, the variations among clients are well minimized. By applying FL on the translated images of all clients, the updated parameters are more consistent across clients, leading to better convergence and better performance. Experimental results validated that the proposed VAFL framework outperforms the current FL framework, especially when dealing with multi-source decentralized small datasets.

We summarize the key differences between our work and the federated domain adaptation work in [8] in the following:

1) Different tasks: The primary goal in [8] was to implement transfer learning from some source clients to the target client under federated learning and improve the performance on the target client. Therefore, the target client is more like a "server," which can access the aggregated gradients from other clients. In contrast, the primary goal of our work is to improve the performance of the global model for classification on all clients (not targeting any specific client), and no client can access the gradients from others.
2) Different approaches for variation reduction: In [8], the cross-client variation problem was addressed through feature-level domain adaptation, while we employ image-level domain adaptation to reduce cross-client variations. One benefit of our framework is that variation reduction can be accomplished offline and completely

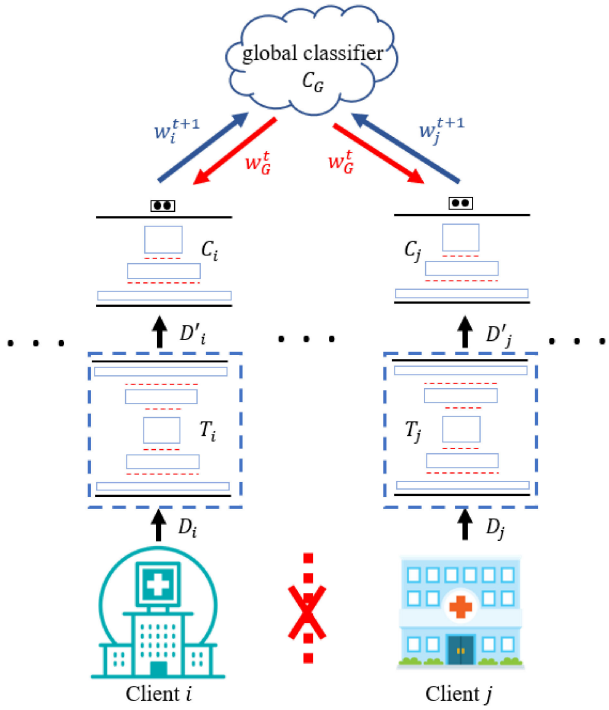

Fig. 2. Overview of the VAFL framework. T_i and T_j are the image-to-image translation modules, and other components are the same as those in the FL framework of Fig. 1.

asynchronous among clients, and thus will not increase communication burdens during federation training.

3) Different data: The framework proposed in [8] was designed and evaluated based on natural image data, while our framework is designed for medical image data. The difference in the requirements on the FL framework between natural image data and medical image data has been discussed before. Besides, the proposed VAFL framework takes into consideration the limitation in the amount of training data, which usually is not a major concern in dealing with natural image data.

The paper is organized as follows. Section II presents details of the proposed VAFL framework. We evaluate the effectiveness of the proposed framework through comparison and analysis of multiple experiments in Section III. Section IV presents discussions and Section V concludes the paper.

II. Framework

The core idea of VAFL is to alleviate the variations among different clients by transforming the raw medical image data of all clients onto a common image space via image-to-image translation, without violating the privacy setting in FL. In contrast to the typical horizontal FL in Fig. 1, each client i performs local model update based on the translated data D'_i (generated by the image-to-image translation module T_i), instead of the raw data D_i as shown in Fig. 2. Given any two clients i and j, as the variations between D'_i and D'_j are effectively minimized through T_i and T_j, the updated parameters w_i and w_j are more consistent,

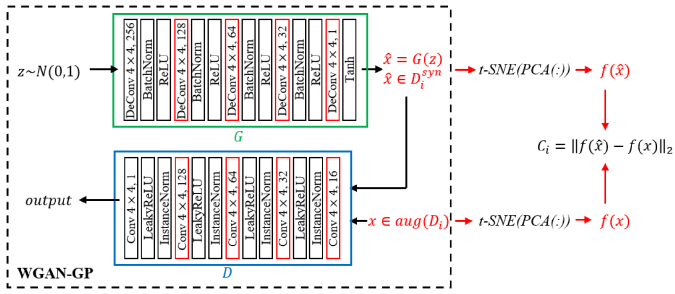

Fig. 3. The workflow of complexity measurement. $aug()$ represents the data augmentation operation (*e.g.*, random non-rigid image deformation by default), and it is used to stabilize the training process for better synthesis.

leading to more stable parameters w_G, better convergence and better performance.

To perform image-to-image translation in VAFL, the first step is to determine the target image space. Given N clients, we select one client k whose training data is of the lowest complexity to define the target image space, and each of the $N-1$ clients translates its raw image data to the target image space. Instead of directly sharing the raw data D_k of the client k with other clients which is not allowed in FL, we propose to share a collection of synthesized data D_{shared} generated by a privacy-preserving image synthesis framework based on D_k. Then, for each client i, based on its raw data D_i and the shared synthesized data D_{shared} representing the target image space, an image-to-image translation model T_i is trained locally to translate the raw data D_i in its original image space to D'_i in the target image space. In this way, the raw data of all clients in VAFL can be translated into a common image space with maximized homogeneity. Details are described in the following.

A. Target Client Selection

To select a suitable client k out of N clients as the target to perform image synthesis for sharing, a complexity measure is constructed to evaluate the data complexity of each client as shown in Fig. 3. For each client i, given its raw data D_i, we first train a Wasserstein generative adversarial network with gradient penalty (WGAN-GP) [19] for certain epochs (*e.g.*, 100 by default) to synthesize a collection of images D_i^{syn} with the same size as D_i, using the following loss functions:

$$\begin{aligned} \mathcal{L}_G &= -\mathbb{E}[D(G(z))], \\ \mathcal{L}_D &= -\mathbb{E}[D(x)] + \mathbb{E}[D(G(z))] \\ &\quad + \lambda \mathbb{E}[(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)^2], \end{aligned} \tag{1}$$

where $\tilde{x} = \alpha \cdot x + (1-\alpha) \cdot G(z)$, $\alpha \sim (U)(0,1)$, and $\lambda = 10$. Then, we perform PCA and t-SNE to extract the features of both D_i and D_i^{syn}, denoted as f_i and f_i^{syn}. The complexity score C_i of D_i is defined as the L2 distance between f_i and f_i^{syn}. As the number of training epochs is fixed, a lower C_i indicates that it is easier to synthesize D_i and the characteristics of D_i are easier to capture, which is helpful for the following image-to-image translation process. Therefore, the client with the lowest complexity score is selected as the target client.

When dealing with continuous data (*i.e.*, clients are dynamically added), given a newly added client, we first calculate its data complexity according to the above process and compare it with the currently selected client. If its complexity is notably lower, the newly added client will be selected as the new target client, and the entire VAFL framework will be re-trained accordingly. Otherwise, an image-to-image translation module would be trained locally for the newly added client and then added for federation training.

B. Privacy-Preserving Image Synthesis and Sampling

Once the target client k is selected according to Section II.A, the next step is to synthesize data based on client k's raw data D_k. Note that the generalization capability of generative adversarial networks is critical for our application for assuring that new images somewhat similar to but also sufficiently distinct from those images in D_k can be successfully synthesized. Similar to supervised training, prevention of over-fitting is the key to generalization, which usually is achieved by either using a large amount of training data, relying on appropriate network architecture, or through regularization approaches. Considering that properly labeled medical image data available for training is often limited and expensive to acquire, we therefore focus more on the regularization approaches. Regularization based on gradient penalty [19] is an effective technique for not only preventing over-fitting but also improving the generalization capability of GAN. More importantly, imposing gradient penalty as a Lipschitz regularization technique has been proven effective in alleviating information leakage of the training data [20], especially in tackling the membership inference attack [21]. As privacy protection is the primary requirement of FL, we therefore choose WGAN-GP, as shown in Fig. 3, as the base model for image synthesis. Following the approach in [22], we add certain well-designed noise to the gradients of the discriminator D during training to further protect the privacy of the raw data, denoted as privacy-preserving WGAN-GP (PPWGAN-GP). Based on the definitions given in [22], the noise is set as $N(0, \sigma_n^2 c_g^2 I)$, where $c_g = 1$ according to gradient penalty and I is the identity matrix. σ_n is defined as

$$\sigma_n = 2q\sqrt{n_d \log \frac{1}{\delta}}/\epsilon, \tag{2}$$

where q is the sampling probability, $n_d = 5$ is the update frequency of the discriminator D, δ is set as $1.0\times^{-5}$, and ϵ is set as 10. Observing from our experimental results, adding noise to gradients would not affect the final synthesis results much. One possible reason is that imposing gradient penalty, to some extent, has a similar effect as introducing additional noise. The proofs of privacy preservation of the deployed PPWGAN-GP have been presented in [20], [22].

Note that the privacy leakage problem of generative adversarial networks usually refers to information leakage from the well-trained generators [21]. Therefore, to avoid any possible privacy leakage, the trained generator is not shared with other clients. Instead, only synthesized images are shared. We further implement an image sampling process to select only a subset of synthesized images D_{shared} for sharing. The selected subset D_{shared} satisfies the following two requirements:

1) The distributions of D_{shared} and D_k are similar in the feature space to benefit the following image-to-image translation process.
2) Every raw image in D_k is sufficiently distant from D_{shared} in the image space.

The above image sampling process is formulated as the following feasibility problem:

$$\begin{aligned} \text{find} \quad & \hat{x} \\ \text{subject to} \quad & \|\hat{x} - x\|_2 > \epsilon \\ & \|\hat{x}_f - x_f\|_2 < \delta \times C \\ & \delta \in (0,1], C = \max_{i,j=0,1,..,|D_k|-1} \|x_f^i - x_f^j\|_2, \end{aligned}$$

where $\hat{x}$ is a synthesized image generated by PPWGAN-GP, $\hat{x}_f$ is the feature vector of $\hat{x}$ in the feature space, $x \in D_k$ is the 1-nearest neighboring image of $\hat{x}$ in the image space, x_f is the feature vector of x in the feature space, ϵ is a threshold of L2 distance in the image space, and δ is a threshold of L2 distance in the feature space. The feature space is defined by performing T-SNE and PCA onto the set of raw and synthesized images,

Algorithm 1: Image Sampling.

input: D_k, raw images of the target client k,
G, the generator of PPWGAN-GP
parameter: $\delta = 0.2$, threshold in feature space,
N, size of D_{shared}
output: D_{shared}, sampled synthesized images

Initialize D_{shared} := {}
while $|D_{shared}| \leq |D_k|$ **do**
  Initialize D_{syn} := {}
  while $|D_{syn}| \leq |D_k|$ **do**
    sample $z \sim p(z), z \in N[0,1]$
    D_{syn}.insert($G(z)$)
  end while
  $F_{syn}, F_k \leftarrow$ T-SNE(PCA($D_{syn} \cup D_k$))
  $C \leftarrow \max \|x_f^i - x_f^j\|_2, \forall x_f^i \in F_k, x_f^j \in F_k$
  for $\hat{x} \in D_{syn}, \hat{x}_f \in F_{syn}$ **do**
    $x \leftarrow \arg\min_x \|x - \hat{x}\|_2, \forall x \in D_k$
    if $\|\hat{x}_f - x_f\|_2 < \delta \times C$ **then**
      Initialize x_{5nn} := {}
      while $|x_{5nn}| \leq 5$ **do**
        $x_{nn} \leftarrow \arg\min_{x^i \in D_k - x_{5nn}} \|x_f - x_f^i\|_2$
        x_{5nn}.insert(x_{nn})
      end while
      $\epsilon \leftarrow \max \|x - x^i\|_2, \forall x^i \in x_{5nn}$
      if $\|\hat{x} - x\|_2 > \epsilon$ **then**
        D_{shared}.insert($\hat{x}$)
      end if
    end if
  end for
end while
return D_{shared}

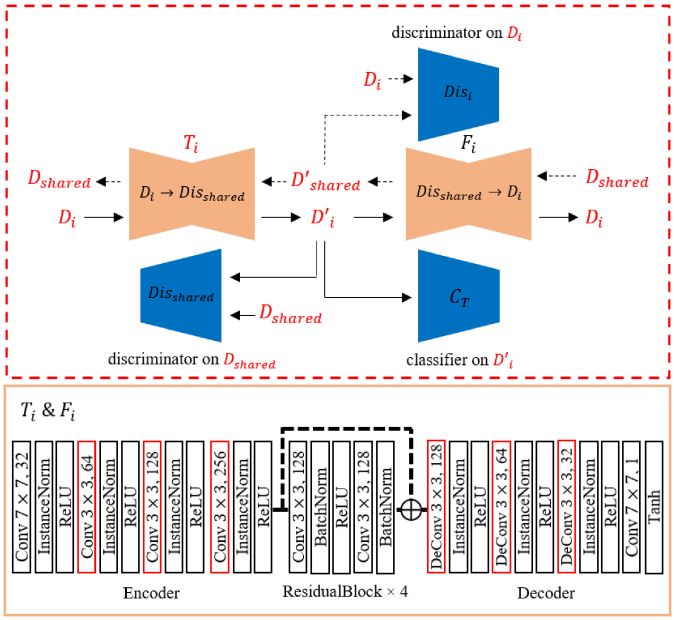

Fig. 4. Overview of the image-to-image translation module.

while the image space is defined as the pixel-wise intensity values. To ensure that the feasibility problem is data-invariant, the value of ϵ is dynamically adjusted by finding the maximum L2 distance in the image space between x and its, say, 5-nearest raw neighboring images in the feature space. Then, given a synthesized image $\hat{x}$, if the L2 distance between $\hat{x}$ and x in the feature space is less than $\delta \times C$ (*i.e.*, satisfying Requirement 1) above) and the L2 distance between $\hat{x}$ and x in the image space is larger than ϵ (*i.e.*, satisfying Requirement 2) above), $\hat{x}$ will be selected for inclusion in D_{shared}. The image sampling process iterates untill the size of D_{shared} is the same as that of D_k. The pseudo code of the image sampling process is summarized in Algorithm 1.

C. Image-to-Image Translation

For each client i, based on its raw training data D_i and the shared synthesized data D_{shared}, CycleGAN [23] can be used to learn the mapping $D_i \rightarrow D_{shared}$ for image-to-image translation. It should be mentioned that both D_i and D_{shared} are augmented via random non-rigid image deformation, to ensure a sufficient amount of training data for high-quality image-to-image translation. It is helpful, especially when the amount of training data is relatively limited. A modified CycleGAN framework is shown in Fig. 4. The transformer T_i aims to translate D_i to D'_i, so that D'_i is similar to D_{shared}. The discriminator Dis_{shared} competes with T_i to differentiate the translated data in D'_i and the real data in D_{shared}. Thus, T_i and Dis_{shared} can be trained by adversarial learning with the following objective function:

$$\mathcal{L}_{adv}(T_i, Dis_{shared}) = \mathbb{E}_{\hat{x}_i \sim D_{shared}}[Dis_{shared}(\hat{x}_i)]^2 + \mathbb{E}_{x_i \sim D_i}[1 - Dis_{shared}(T_i(x_i))]^2, \quad (3)$$

where T_i aims to minimize the objective function and Dis_{shared} is to maximize the objective function.

To preserve the original contents of D_i, a reconstruction loss is defined as:

$$\mathcal{L}_{cyc}(T_i, F_i) = \mathbb{E}_{x_i \sim D_i} \left\| F_i(T_i(x_i)) - x_i \right\|_1. \quad (4)$$

One problem of this reconstruction loss is that it treats all the pixels in x_i equally. For a medical image, usually only a small portion of the image is crucial for the diagnosis/classification task. In other words, the majority of pixels in the image is somewhat irrelevant for classification. Thus, with this reconstruction loss, those crucial regions can not be properly emphasized. As a result, the translated medical image may lose useful information for classification. To better preserve those crucial regions, we add a classifier C_T trained by

$$\mathcal{L}_{classify}(T_i, C_T) = \mathbb{E}_{(x_i, y_i) \sim D_i}[y_i - C_T(T_i(x_i))]^2. \quad (5)$$

As the classification task of C_T is to ensure that the medical image after translation is available for classification/diagnosis, imposing this classification loss can better preserve those crucial regions. In addition, the identity loss is defined as

$$\mathcal{L}_{iden}(T_i) = \mathbb{E}_{\hat{x}_i \sim D_{shared}} \| T_i(\hat{x}_i) - \hat{x}_i \|_1. \quad (6)$$

Then, the overall loss of the transformer T_i becomes:

$$\mathcal{L}(T_i; C_T, Dis_{shared}, F_i) = \mathcal{L}_{adv} + 10 \cdot \mathcal{L}_{cyc} + 5 \cdot \mathcal{L}_{iden} + \mathcal{L}_{classify}, \quad (7)$$

F_i and Dis_i are trained in a similar way, except for the classification loss. After training, parameters of the image-to-image translation module are frozen, and medical image data of each client will undergo the transformation of T_i before being used for FL. It should be pointed out that image-to-image translation does not have to be "perfect," as FL itself can handle a reasonable degree of variations among clients. It is because the server in FL works more in a centralized learning way (except for only the parameters rather than the raw data of clients are available to the server) instead of in a transfer learning way from one client to another.

III. Evaluation

To conduct a case study, we apply the proposed VAFL framework for the classification of clinically significant prostate cancer (CS PCa) based on multi-source decentralized apparent diffusion coefficient (ADC) image data. As one of the most common cancers in men, it has been estimated that, in 2020, around 191,930 new cases of PCa would be diagnosed and nearly 33,330 deaths caused by PCa would occur in the United States [24]. Among all cases, nearly 90% PCa patients are low-risk, *i.e.*, whose Gleason scores are lower than 7, and only need active surveillance. Comparatively, those patients with clinically significant (CS) PCa, *i.e.*, whose Gleason scores are equal to or greater than 7, would need timely diagnosis and proper treatment [25]. Therefore, accurate identification of CS PCa can significantly increase the survival rate of patients and reduce the risk of overtreatment for patients with indolent PCa, which is highly valuable in practice.

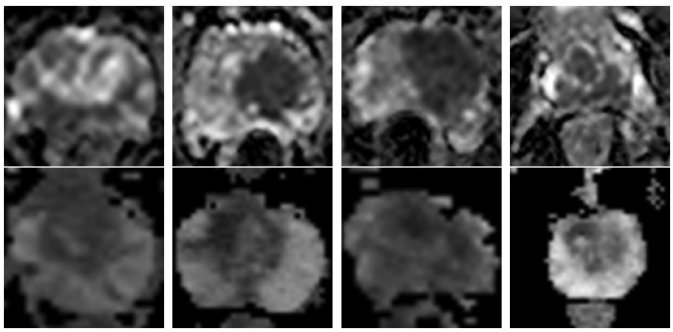

Fig. 5. Exemplar ADC images from the two datasets. Row 1: The locally collected dataset. Row 2: The PROSTATEx dataset [26]–[28].

A. Datasets and Metrics

ADC images from two datasets have been used for evaluation, and some exemplar images are shown in Fig. 5. The first dataset consisting of 135 patients, named LocalPCa, was locally collected. The second dataset was the training set from the PROSTATEx challenge [26]–[28], of which 64 CS PCa patients and 124 nonCS PCa patients were selected based on the quality of pixel-wise annotations by our experts.

For evaluation, the performance of image-level CS PCa vs. nonCS PCa classification is evaluated using the area under curve (AUC) and the balanced classification accuracy (Acc) calculated by using a fixed threshold of 0.5. It is because the number of images per patient varies among patients.

B. Network Architectures

To provide stable qualitative results, inspired by [29], we construct a specially-designed classier used in different learning frameworks as shown in Fig. 6. It is trained by

$$\mathcal{L}_C(x,y) = \ell(p,y) + 0.1 * \ell_1(p_1,y) + 0.1 * \ell_2(p_2,y) + \ell(CAM_0, CAM_1), \quad (8)$$

where p is $(p_1+p_2)/2$, ℓ, ℓ_1 and ℓ_2 are the cross entropy loss functions, and $\ell(CAM_0, CAM_1)$ is defined as

$$\ell(CAM_0, CAM_1) = |CAM_0 \cup CAM_1| + y * |CAM_0| + (1-y) * |CAM_1|, \quad (9)$$

where CAM_0 and CAM_1 are the normalized class activation maps. $\ell(CAM_0, CAM_1)$ enforces additional constraints to prevent overfitting and strengthen the relationship between the predictions and the corresponding class activation maps. Observing the class activation maps produced by different learning frameworks, we can qualitatively evaluate the performance.

C. Implementation Details

All deep learning frameworks/models were trained using Adam optimizer with an initial learning rate of 1e-4 and the batch size of 24. All methods were implemented within the PyTorch framework and trained on Nvidia GeForce Titan Xp GPUs for 100 epochs.

The settings of all FL frameworks are the same, including the classifier architecture, the aggregation function (*i.e.*, FedAvg [5]), etc. Similar to [8], [9], we simulate the federated systems on a single machine (as data communication is not the main focus of this paper and is not a bottleneck for our experiments), and adopt synchronous federated optimization.

D. Learning Frameworks

To evaluate the effectiveness of VAFL, several learning frameworks consisting of 2, 4, and 8 clients have been constructed for comparison, including:

1) *Local learning (LL):* For each client, a local deep learning model is trained by its training images. Given N clients, totally N deep learning models were trained and tested separately for evaluation.
2) *Centralized learning (CL):* The training data of all clients is fused to train a global model, assuming privacy is not a concern. Given N clients, only one global model was trained and tested on the concatenated testing data of all clients for evaluation.
3) *Federated learning (FL):* The parameters of the global model are updated based on clients' raw image data and employing FedAvg for aggregation. Given N clients, only one global model was trained and tested on the testing data of all clients for evaluation.
4) *Variation-aware federated learning (VAFL):* The parameters of the global model are updated based on clients' translated image data and employing FedAvg for aggregation. Given N clients, only one global model is trained based on the aforementioned process which is tested on the testing data of all clients for evaluation.

By comparing the results of VAFL with FL, we can assess how well the proposed VAFL can address the cross-client variation problem. In addition, by comparing the results of LL with FL and VAFL under different settings, we can identify the situations that VAFL outperforms LL and thus benefits clients from collaboration through federation training. To some extent, CL can be regarded as the "upper bound" of FL, as CL can access the training data of all clients to reduce the cross-client variations. Through comparison with CL, we can evaluate the effectiveness of VAFL in addressing the cross-client variation problem under the FL setting.

E. Intermediate Results of Image Synthesis

Exemplar randomly synthesized images are shown in Fig. 7. In 2-client VAFL, based on the calculation process in Section II.A, the client *C-2* (*i.e.*, PROSTATEx) is selected as the target for image synthesis. It is consistent with our observation that images of *C-1* (*i.e.*, LocalPCa) contain more details and thus are more challenging for synthesis, leading to a higher complexity score. From the synthesized images, we find that the appearance characteristics of the raw images in *C-2* are well captured. The distributions of the synthesized images being consistent with that of the raw images validates the effectiveness of the privacy-preserving image synthesis framework PPWGAN-GP.

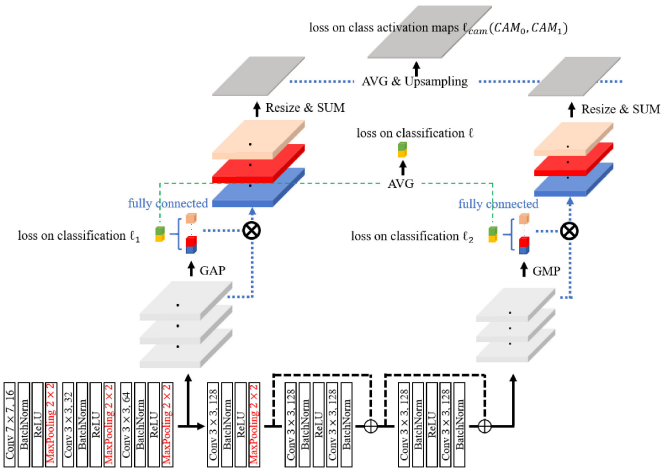

Fig. 6. Overview of the network architecture of the deployed classifiers.

With the increase of the number of clients, the amount of training data of each client is further reduced. In 4-client VAFL, though the amount of training data after augmentation of the target client *C-3* is only 50% of that of *C-2* in 2-client VAFL, the adopted PPWGAN-GP can stably generate high-quality synthesized images. Similar results can be observed in 8-client VAFL, where the amount of training data of the target client *C-5* is only 25% of that of *C-2* in 2-client VAFL. Based on the results in the image space and the feature space, we find that PPWGAN-GP can effectively capture the characteristics of the training data and synthesize high-quality images, which is crucial for image-to-image translation.

F. Intermediate Results of Image-to-Image Translation

Some exemplar image-to-image translation results of *C-1* in 2-client, 4-client, and 8-client VAFL frameworks are shown in Fig. 7. As image-to-image translation is trained based on the augmented images of each client via random non-rigid image deformation, the total amount of training data can ensure the stability of the modified CycleGAN, leading to consistent results. Given the same raw images, though the corresponding translated images are slightly different, the appearance characteristics are almost the same, regardless of the amount of available training data. Comparing the images before and after translation, the less important background pixels are effectively suppressed and most of the cancerous regions can be preserved for classification. The image-to-image translation results further demonstrate the effectiveness of PPWGAN-GP for image synthesis.

G. Results of 2-Client Frameworks

1) Data Preparation: Both LocalPCa and PROSTATEx are separately and randomly divided into the 5 subsets. In each subset for training, both CS PCa images and nonCS PCa images are augmented so that the ratio is relatively balanced. Augmentation includes random non-rigid image deformation and random flip. In the 2-client experiment, we assume LocalPCa is client *C-1*'s data and PROSTATEx is *C-2*'s data. More details can be found in Table I.

TABLE I
DATA OF 2-CLIENT LEARNING FRAMEWORKS (SPLIT RATIO: 80%)

Client		patients	raw images	augmented images
C-1	Train	111	680	9790
	Test	24	130	-
C-2	Train	152	1008	7888
	Test	36	255	-

2) Qualitative Results: According to the qualitative results in Fig. 8, when the amount of clients' data is sufficient to train domain-invariant features, the class activation maps generated by LL, CL, FL, and VAFL can effectively locate most cancerous regions. In general, the class activation maps produced by VAFL are much "cleaner" and contain fewer false positives than those of other learning frameworks.

With the decrease of the split ratio, there would exist more false positives in the class activation maps produced by both CL and FL. FL fails to capture the cancerous regions in some cases, leading to incorrect classification. It is due to the bad convergence caused by the cross-client variations. Observing the class activation maps obtained by VAFL, the most important cancerous regions are well detected, leading to correct classification, especially for *C-1*.

3) Quantitative Results: Quantitative results of a set of 2-client experiments are listed in Table II. For each client, four different patient-wise splits of the client's dataset are experimented for training and testing. Specifically, 20%, 40%, 60%, and 80% respectively of the client's data were used for training and the rest of its data not used for training were used for testing (*i.e.*, 80%, 60%, 40%, and 20% respectively).

Let's first discuss the case when 80% of the clients' data is used for training (thus the rest 20% used for testing), which represent the case where the image data of each client is sufficient for training its local deep learning model, and thus mimicking the situation when FL is least beneficial. As shown in the rightmost two columns of Table II (labelled as 80% split ratio of Train/Total), compared to LL, both FL and VAFL fail to further improve the overall performance. Compared to CL, though the performance of VAFL is quite close to CL, it still incurs 0.81% drop in Acc and 1.07% drop in AUC. It indicates that when data of each client is sufficient for training, LL is the best strategy, as a client's testing data often shares a similar distribution as its training data.

4) Discussions on Split Ratio: To explore the situations when deploying VAFL can outperform LL, we conduct additional experiments by changing the split ratio of each client. By decreasing the split ratio from 80%-20%, to 60%-40%, 40%-60% and 20%-80% respectively, the amount of image data of each client used for training is reduced proportionally. In this way, we can evaluate the effectiveness of VAFL especially in learning from multi-source decentralized small datasets.

Quantitative results of these 2-client experiments with respect to these 4 split ratios are listed in Table II. With the decrease of the split ratio, the inter-patient variation problem between

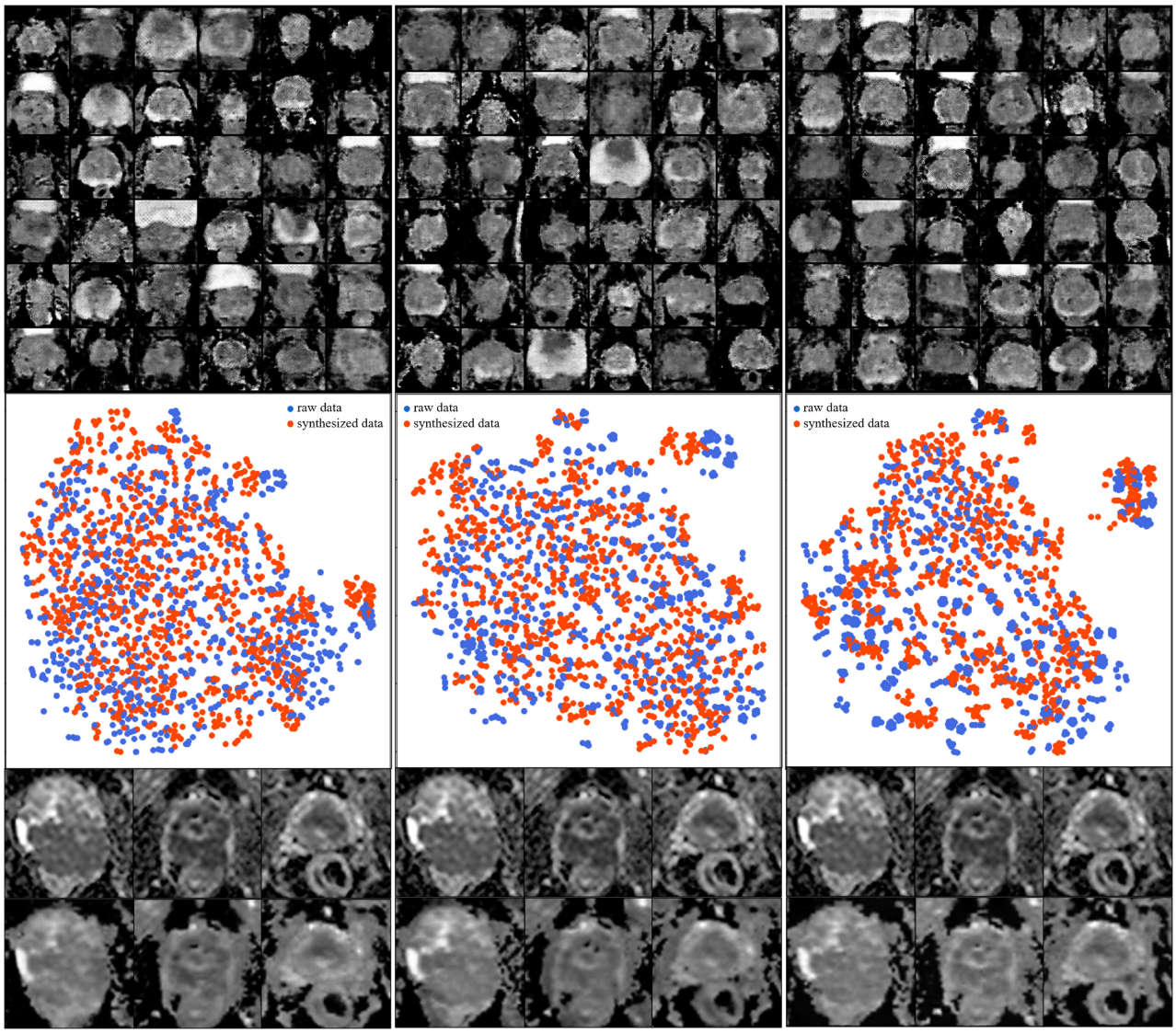

Fig. 7. Intermediate results of image synthesis and image-to-image translation. Row 1: The synthesized images of 2-client, 4-client, and 8-client VAFL respectively. Row 2: The distributions in feature space between the synthesized images and the augmented raw images of 2-client, 4-client, and 8-client VAFL respectively. Row 3: The raw images and the corresponding translated images of 2-client, 4-client, and 8-client VAFL respectively. To plot distributions, we randomly selected 1000 raw images (after augmentation) and 1000 synthesized images for visualization.

TABLE II
COMPARISON RESULTS OF DIFFERENT 2-CLIENT LEARNING EXPERIMENTS WITH CHANGING SPLIT RATIO

Framework	Client ID	Split Ratio (Train/Total)							
		20%		40%		60%		80%	
		Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
LL	*C-1*	0.8825	0.9322	0.9016	0.9608	0.9535	0.9944	0.9787	0.9993
	C-2	0.8849	0.9145	0.9026	0.9164	0.9282	0.9407	0.9836	0.9994
CL	*C-1*	0.9084	0.9384	0.9224	0.9704	0.9698	0.9977	0.9949	0.9999
	C-2	0.8790	0.9186	0.9116	0.9604	0.9734	0.9983	0.9675	0.9980
	Overall[1]	0.8938	0.9293	0.9173	0.9658	0.9716	0.9981	0.9803	0.9982
FL	*C-1*	0.8280	0.9262	0.9002	0.9684	0.9155	0.9692	0.9587	0.9965
	C-2	0.8480	0.9018	0.8782	0.9424	0.9559	0.9882	0.9630	0.9996
	Overall[1]	0.8288	0.8995	0.8844	0.9514	0.9389	0.9806	0.9655	0.9981
VAFL	*C-1*	0.9278	0.9699	0.9271	0.9735	0.9429	0.9716	0.9898	0.9959
	C-2	0.8587	0.9108	0.8836	0.9346	0.9539	0.9894	0.9580	0.9818
	Overall[1]	0.8858	0.9375	0.9005	0.9517	0.9488	0.9823	0.9722	0.9875

[1]The overall scores of Acc and AUC are calculated by concatenating the predictions and labels from *C-1* and *C-2*.

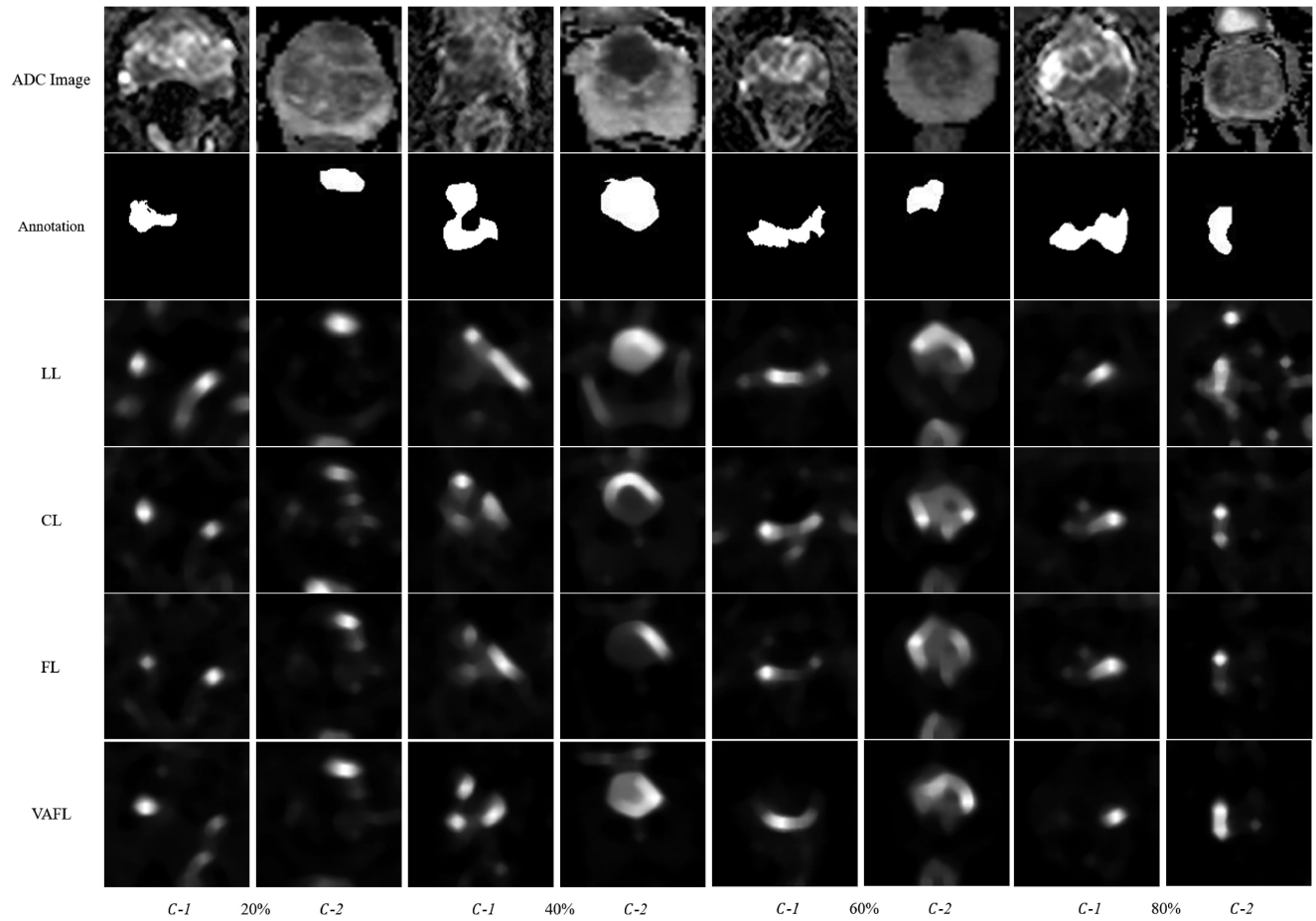

Fig. 8. Exemplar results of 2-client learning frameworks. Rows 1 to 6: The raw ADC images, the annotated cancerous regions, the class activation maps generated by LL, CL, FL, and VAFL respectively. Here, the class activation maps are normalized into intensity maps, where the intensity value of each pixel represents the confidence for prediction. Pixels with higher intensity values contribute more to final predictions.

training and testing becomes more severe. From the results, we find that VAFL can achieve a performance quite close to that of CL, which validates the effectiveness of VAFL in alleviating cross-client variations. Comparatively, when decreasing the split ratio, the performance gap between FL and CL is enlarged, due to the cross-client variation problem.

We further examine the performance gains of *C-1* and *C-2* separately under different split ratios as shown in Fig. 9. For *C-1*, the performance of FL is always worse than LL, and the gap between FL and LL is unstable largely depending on the cross-client variations. Comparatively, VAFL steadily outperforms LL, and the gap between VAFL and LL gradually increases along with the decrease of the split ratio. Specially, when the split ratio is further decreased to 40% and 20%, VAFL would outperform CL, which confirms the effectiveness of VAFL for variation reduction.

For *C-2*, the trends are quite different compared to *C-1*. Since the inter-patient variation problem between the training data and the testing data of *C-2* is more severe than that of *C-1*, the testing performance highly depends on the inter-patient variations. For most cases, CL, FL, and VAFL fail to improve the performance of LL. It is because that CL, FL, and VAFL are trained globally to achieve the best overall performance across clients, and the training process can be biased to some specific clients depending on the training data. It explains why the curves of CL, FL, and VAFL are very similar. In the meantime, for most cases, VAFL effectively outperforms FL and the performance gap between VAFL and FL increases with the decrease of the split ratio. It validates the effectiveness of VAFL compared to FL, especially when the training data is relatively limited.

H. Results of 4-Client Frameworks

1) Data Preparation: Both LocalPCa and PROSTATEx are separately and randomly patient-wisely partitioned into 2 subsets, resulting in a total of 4 sets of data, and we assume in our experiment that each of the 4 sets is the data of each of the 4 clients. For each client, 80% of its image data is used for training and the rest 20% is used for testing. The same augmentation approaches were utilized for augmenting the data. We assume that the two subsets of LocalPCa were the image data of clients *C-1* and *C-2* respectively, while the two subsets of PROSTATEx were the image data of *C-3* and *C-4*. More details are in Table III.

2) Qualitative Results: According to the qualitative results in Fig. 10, the class activation maps generated by LL in some

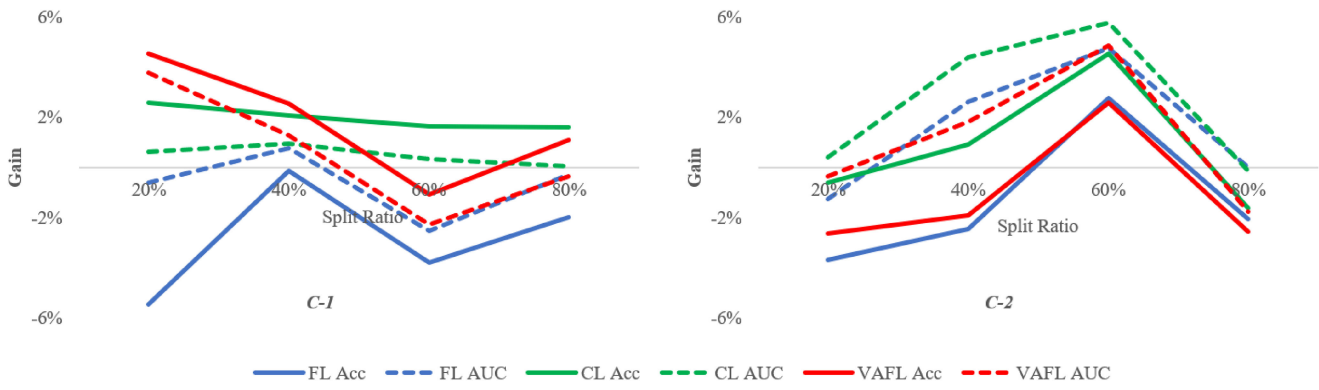

Fig. 9. Gains achieved by FL and VAFL compared to LL with changing split ratio. The vertical axis is the performance difference, in terms of Acc or AUC, between FL and LL or between VAFL and LL. A positive (negative) gain means FL or VAFL achieves better (worse) performance than LL.

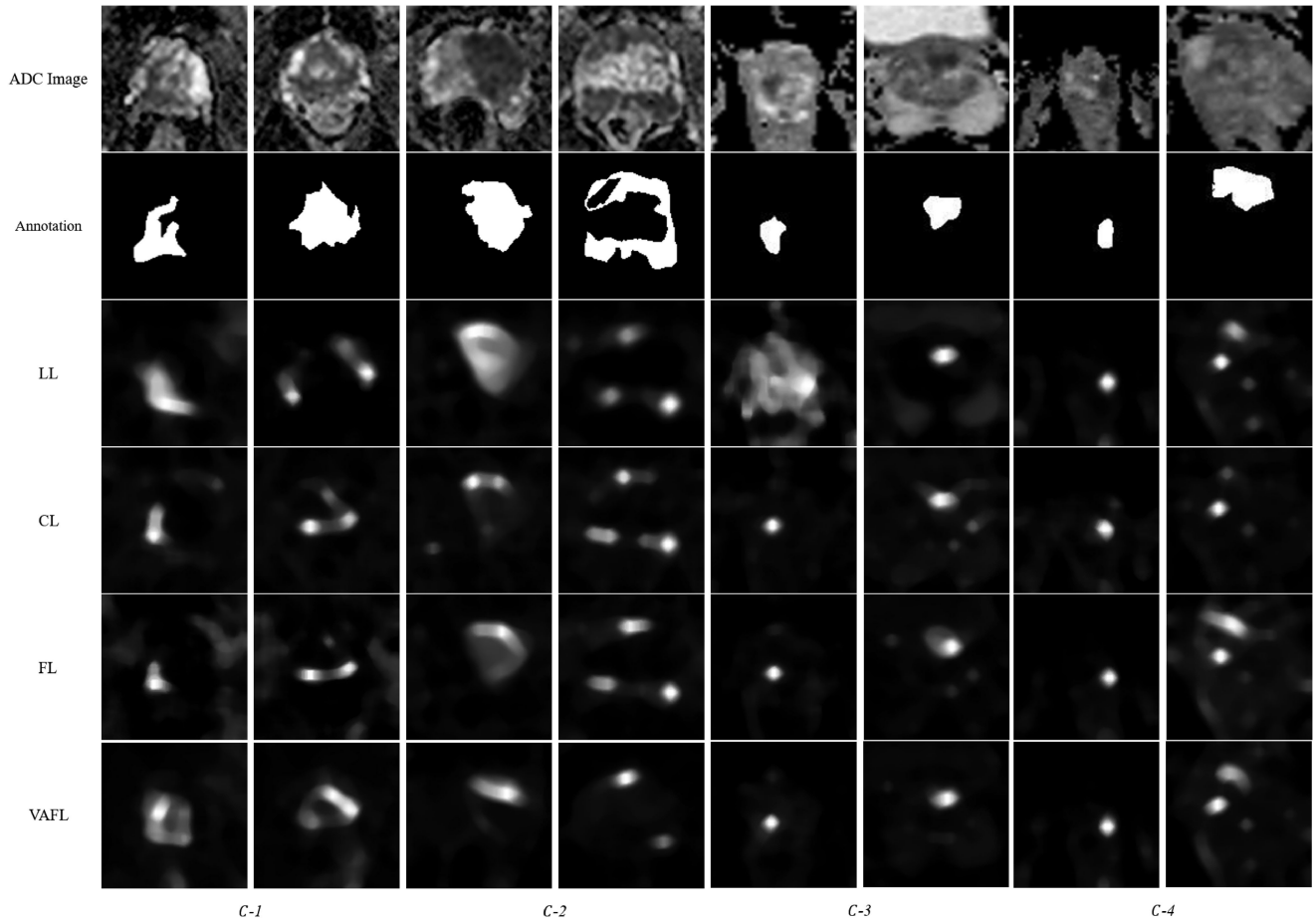

Fig. 10. Exemplar results of 4-client learning frameworks. Rows 1 to 6: The raw ADC images, the annotated cancerous regions, the class activation maps generated by LL, CL, FL, and VAFL respectively. Here, the class activation maps are normalized into intensity maps, where the intensity value of each pixel represents the confidence for prediction. Pixels with higher intensity values contribute more to final predictions.

cases contain more false positives and fail to detect the cancerous regions, leading to incorrect classification. Comparatively, CL, FL, and VAFL can produce better class activation maps. It is because the classifiers trained by LL can not perfectly handle the inter-patient variation problem, due to the limited amount of training data. Since CL, FL, and VAFL are trained by the data of all clients, they have a better chance to learn more generalizable features to reduce the inter-patient variations. Compared to CL and VAFL, the class activation maps produced by FL contain more false positives, especially for *C-1*. It should be pointed out that, though in some cases the cancerous regions detected by VAFL is not complete, it is sufficient for correct classification.

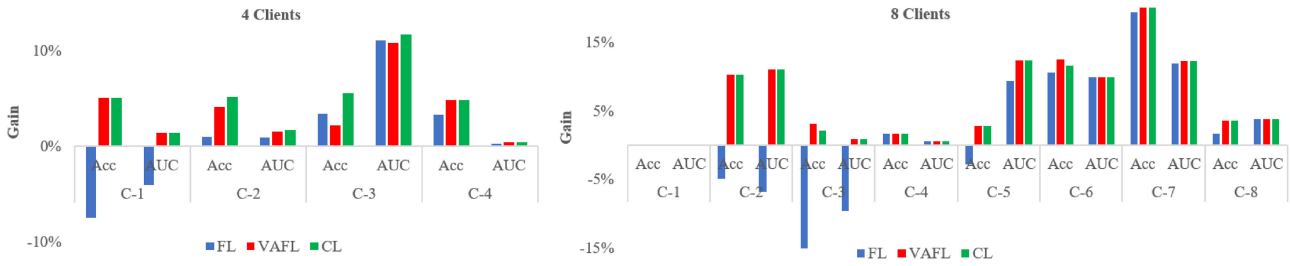

Fig. 11. Gains achieved by FL and VAFL compared to LL. The vertical axis is the performance difference, in terms of Acc or AUC, between FL and LL or between VAFL and LL. A positive (negative) gain means FL or VAFL achieves better (worse) performance than LL.

TABLE III
DATA OF 4-CLIENT LEARNING FRAMEWORKS (SPLIT RATIO: 80%)

Client		patients	raw images	augmented images
C-1	Train	56	353	5083
	Test	12	69	-
C-2	Train	55	327	4707
	Test	12	61	-
C-3	Train	76	501	3913
	Test	18	133	-
C-4	Train	76	507	3975
	Test	18	122	-

The incompleteness mainly is due to the lack of pixel-wise annotations for guidance during image-to-image translation.

3) Quantitative Results: The results of the 4-client experiments are listed in Table IV. Observing the results of LL, the performance is significantly limited by the inter-patient variations between the training and testing data, due to the reduced amount of training data. When using FL, though it slightly improves the performance of *C-2*, *C-3*, and *C-4*, it suffers from significant performance degradation of *C-1*. In terms of the overall Acc and AUC, FL suffers 4.88% decrease in Acc and 1.48% decrease in AUC compared to CL. Comparatively, VAFL can improve the performance of all clients, approaching to the performance of CL. In Table IV, we find that the Acc and AUC scores of some clients in CL and VAFL are 1.0000. It is because the amount of testing data of each client is further reduced in the 4-client setting compared to the 2-client setting. Comparing the overall performance of both CL and VAFL in Table IV and Table II (when the split ratio is 80%), we find that their overall Acc and AUC scores are consistent, which validates the correctness of those quantitative results.

The performance gains achieved by different learning frameworks compared to LL are illustrated in Fig. 11. It should be pointed out that, with the increase of the number of clients in VAFL, as more clients would have transformers for image-to-image translation, the variations among the training data of all clients are effectively reduced, resulting in better performance compared to that of the 2-client VAFL with the split ratio as 80%. The performance gains of CL and VAFL being similar demonstrates the effectiveness of the proposed framework in variation reduction.

I. Results of 8-Client Frameworks

1) Data Preparation: Both LocalPCa and PROSTATEx are separately and randomly patient-wisely partitioned into 4 subsets, resulting in a total of 8 sets of data, and we assume that each of the 8 sets is the data of each of the 8 clients. For each client, 80% of its image data is used for training and the rest 20% is used for testing. The same augmentation approaches were utilized for augmentation. We assume that the 4 subsets of LocalPCa were the image data of clients *C-1*, *C-2*, *C-3*, and *C-4* respectively, while the 4 subsets of PROSTATEx were the image data of clients *C-5*, *C-6*, *C-7*, and *C-8*. More details can be found in Table V.

2) Qualitative Results: Qualitative results of different learning frameworks are shown in Fig. 12. Similar to the qualitative results of the 4-client setting in Fig. 10, LL in some cases fails to detect the cancerous regions, leading to incorrect classification. Though the general quality of the class activation maps produced by FL is better than that of LL, more false positives can be found compared to CL and VAFL. Usually having more false positives in class activation maps indicates poorer transferability of learned features. Comparing VAFL with CL, we find that their produced class activation maps are quite similar, and in some cases, the class activation maps generated by VAFL are even better. It indicates that VAFL can outperform CL, especially with the increase of the number of clients, as more training data would be translated onto the same image space through transformers.

3) Quantitative Results: The results of the 8-client experiments are listed in Table IV. With the number of clients increased to 8 while the total amount of data remains the same, the amount of image data of each client is reduced accordingly. As a result, when training through LL, the performance of some clients gets worse, mainly due to that the limited training data can hardly handle the inter-patient variation problem when testing. It explains why the performance of clients varies significantly, even though their data is from the same source (*e.g.*, the data of *C-1*, *C-2*, *C-3*, and *C-4* is from LocalPCa). When it comes to FL, though some clients can be improved, the performance of FL is quite unstable. Compared to CL, FL's overall performance degrades by 5% in Acc and 2.93% in AUC. In contrast, VAFL effectively improves the performance of almost all clients, achieving the same performance in Acc and 0.13% increase in AUC compared to CL. Similarly, the phenomenon that more clients in CL and VAFL get 1.0000 in terms of Acc and AUC

TABLE IV
COMPARISON RESULTS OF DIFFERENT LEARNING FRAMEWORKS WITH INCREASING NUMBER OF CLIENTS

Framework	Metric	Client ID								Overall[1]
		C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	
4 Clients										
LL	Acc	0.9500	0.9388	0.9118	0.9524					-
	AUC	0.9867	0.9830	0.8822	0.9958					-
CL	Acc	1.0000	0.9898	0.9667	1.0000					0.9863
	AUC	1.0000	1.0000	0.9984	1.0000					0.9986
FL	Acc	0.8750	0.9481	0.9451	0.9851					0.9375
	AUC	0.9459	0.9915	0.9922	0.9986					0.9838
VAFL	Acc	1.0000	0.9796	0.9333	1.0000					0.9726
	AUC	1.0000	0.9983	0.9896	1.0000					0.9942
8 Clients										
LL	Acc	1.0000	0.8974	0.9167	0.9833	0.9167	0.8656	0.7857	0.9643	-
	AUC	1.0000	0.8895	0.9912	0.9944	0.8744	0.9010	0.8776	0.9629	-
CL	Acc	1.0000	1.0000	0.9737	1.0000	0.9444	0.9811	1.0000	1.0000	0.9830
	AUC	1.0000	1.0000	1.0000	1.0000	0.9978	1.0000	1.0000	1.0000	0.9963
FL	Acc	1.0000	0.8474	0.7588	1.0000	0.8889	0.9717	0.9796	0.9808	0.9330
	AUC	1.0000	0.8211	0.8947	1.0000	0.9678	1.0000	0.9971	1.0000	0.9670
VAFL	Acc	1.0000	1.0000	0.9474	1.0000	0.9444	0.9906	1.0000	1.0000	0.9830
	AUC	1.0000	1.0000	1.0000	1.0000	0.9978	1.0000	1.0000	1.0000	0.9976

[1]The overall scores of Acc and AUC are calculated by concatenating the predictions and labels from all clients.

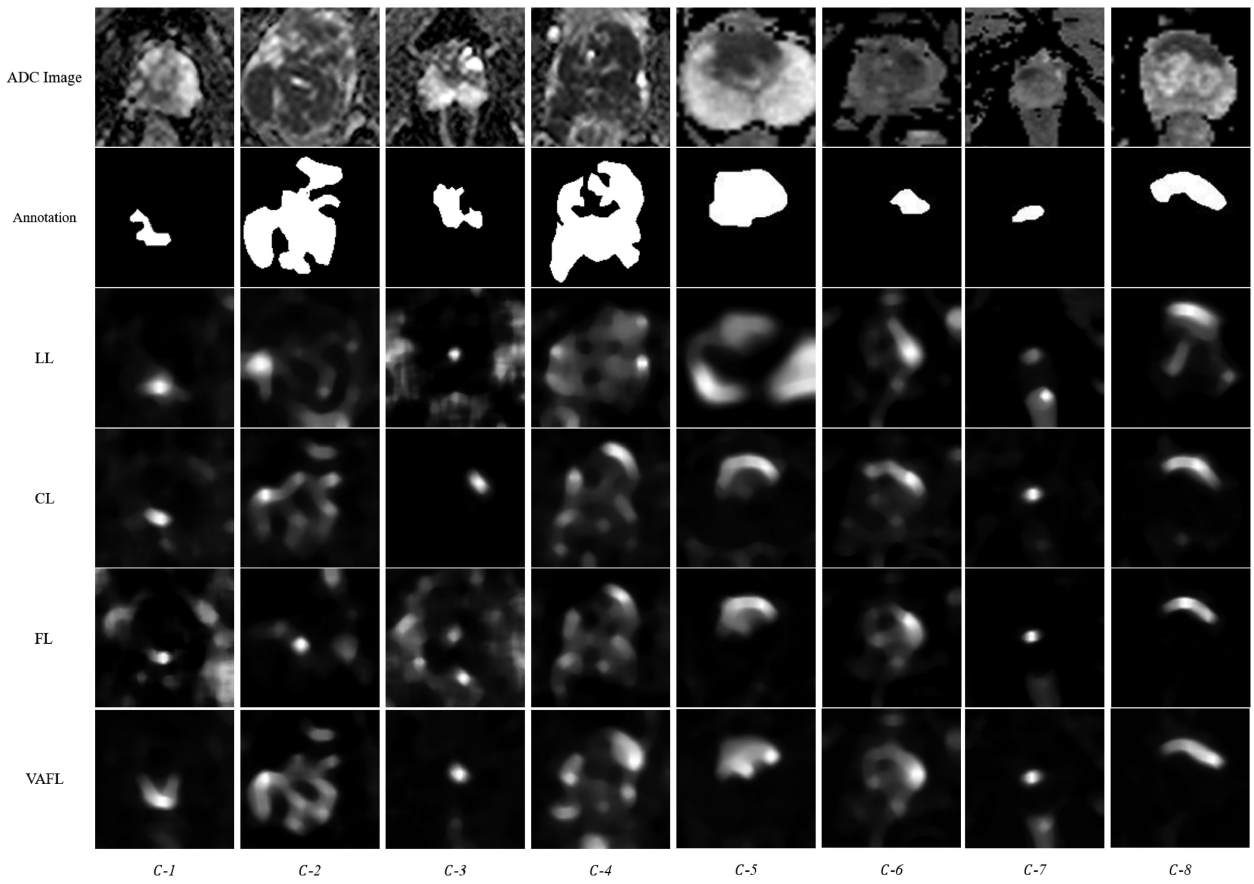

Fig. 12. Exemplar results of 8-client learning frameworks. Rows 1 to 6: The raw ADC images, the annotated cancerous regions, the class activation maps generated by LL, CL, FL, and VAFL respectively. Here, the class activation maps are normalized into intensity maps, where the intensity value of each pixel represents the confidence for prediction. Pixels with higher intensity values contribute more to final predictions.

TABLE V
DATA OF 8-CLIENT LEARNING FRAMEWORKS (SPLIT RATIO: 80%)

Client		patients	raw images	augmented images
C-1	Train	28	199	2789
	Test	6	40	-
C-2	Train	28	154	2294
	Test	6	29	-
C-3	Train	28	138	2038
	Test	6	25	-
C-4	Train	28	189	2669
	Test	6	36	-
C-5	Train	38	253	1953
	Test	9	68	-
C-6	Train	38	248	1960
	Test	9	65	-
C-7	Train	38	257	2037
	Test	9	56	-
C-8	Train	38	250	1938
	Test	9	66	-

scores is due to the reduced amount of testing data of each client. Furthermore, comparing the quantitative results of VAFL with 2, 4, and 8 clients, we find that VAFL achieves very stable performance and can further improve its performance with the increase of the number of clients. The performance gains/losses achieved by CL, FL, and VAFL compared to LL are illustrated in Fig. 11. It clearly shows that VAFL can significantly improve the performance of FL and is valuable to individual clients than LL especially when the amount of data available for training is relatively limited.

IV. DISCUSSION

1) Generalization of VAFL: One benefit of using image-to-image translation to address the cross-client variation problem is that the variation reduction process becomes independent of federation training. As the training of clients' transformers is prior to the training of the global classifier in federated learning, any classifier architecture can be applied to the translated data. Therefore, in the proposed VAFL framework, the classifier can be replaced by other deep learning architectures without affecting the variation reduction task.

2) Computational Cost: Compared to FL, the computational cost of VAFL largely depends on the image synthesis step, which is done by PPWGAN-GP, and the image-to-image translation step, which is done by a modified CycleGAN. Note that these two steps are performed offline prior to federation training. During federation training, the only additional computational cost introduced by VAFL is image-to-image translation through clients' transformers. As the transformers are for inference only and do not require parameter update, the cost can be largely ignored. In addition, the computational cost of VAFL is proportional to the amount of clients' training data. According to the aforementioned experiments, as VAFL is most effective when the clients' training data is relatively limited, the computational cost of VAFL should not be a major concern.

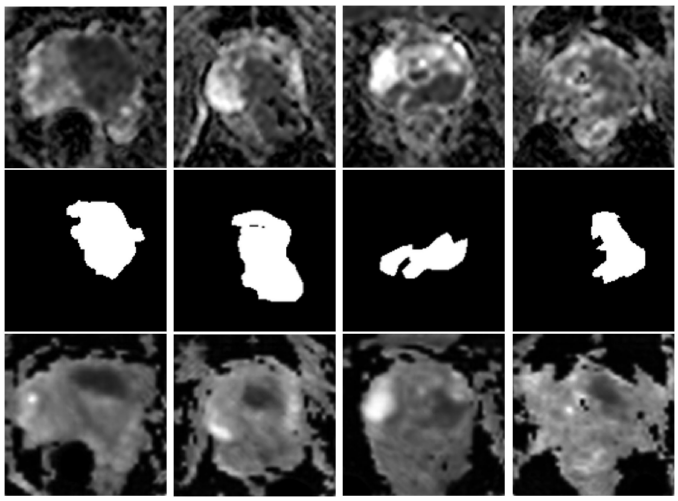

Fig. 13. Exemplar results of image-to-image translation. Rows 1 to 3: The raw ADC images, the annotated cancerous regions, and the corresponding transformed ADC images.

3) Extension to Segmentation: We plan to extend the proposed VAFL framework for medical image segmentation as our future work. In contrast to classification, two unique fundamental problems need to be addressed for medical image segmentation:

1) The inter-observer problem: Given a medical image, it is common that pixel-wise manual annotations marked by different human observers could be different [30], [31]. With multiple clients whose images were annotated by their physicians, how to alleviate the variations in manual annotations of clients' images is crucial for training.
2) Incomplete image-to-image translation: One potential problem of the current image-to-image translation module is the distortion problem of the cancerous regions. Observing the image-to-image translation results in Fig. 13, there exist distortions to the cancerous regions after translation. It explains why for certain cases the detected cancerous regions in the class activation maps produced by VAFL are slightly smaller than those detected by other learning frameworks. Such distortions do not present a serious problem for the classification task; however, they will affect the training process for the segmentation task, since the cancerous regions after translation and the corresponding manual annotations are not very well matched. Therefore, preserving complete cancerous regions during image-to-image translation needs to be addressed before the framework can be successfully extended for the segmentation task.

V. CONCLUSION

Federated learning (FL) has been proven effective in training machine learning models with multi-source decentralized natural image data with privacy preservation. However, the cross-client variation problem which is quite common among multi-source medical image data, has not been properly addressed.

In this paper, we propose a variation-aware federated learning (VAFL) framework which addresses the cross-client variation problem among medical image data. In VAFL, each client first translates its raw training images into a common image space before using them to update the global model. Translating each client's raw training images into a shared image space effectively minimizes the variations among images from different clients. Such image-to-image translation without violating data privacy is achieved by synthesizing images through a privacy-preserving generative adversarial network based on one selected client's training images and only sharing a subset of such synthesized images, not raw images, with all other clients. Experimental results on automated classification of clinically significant prostate cancer from multi-source decentralized ADC images show that the proposed VAFL framework can significantly improve the performance of the current FL framework. While in this paper the assessment was mainly based on prostate cancer data and a classification task, we believe the VAFL framework is applicable to several other types of images and clinical applications.

Acknowledgment

The authors would like to thank the sponsorship of Research Grants Council of Hong Kong and HKUST-WeBank Joint Laboratory.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2017, pp. 1273–1282.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, 2019.

[3] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. ACM Special Int. Group Secur., Audit Control Conf. Comput. Commun. Secur.*, 2015, pp. 1310–1321.

[4] V. Smith, C. -K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 4424–4434.

[5] S. Hardy *et al.*, "Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption," 2017, *arXiv:1711.10677*. [Online]. Available: https://arxiv.org/abs/1711.10677

[6] R. Nock *et al.*, "Entity resolution and federated learning get a federated resolution," 2018, *arXiv:1803.04035*. [Online]. Available: https://arxiv.org/abs/1803.04035

[7] S. Feng and H. Yu, "Multi-participant multi-class vertical federated learning," 2020, *arXiv:2001.11154*. [Online]. Available: http://arxiv.org/abs/2001.11154

[8] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, "Federated adversarial domain adaptation," 2019, *arXiv:1911.02054*. [Online]. Available: http://arxiv.org/abs/1911.02054

[9] C. Ju, D. Gao, R. Mane, B. Tan, Y. Liu, and C. Guan, "Federated transfer learning for EEG signal classification," in *Proc. Eng. Med. Biol. Soc.*, 2020, pp. 3040–3045.

[10] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[11] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: Toward a full-field digital mammographic database," *Acad. Radiol.* vol. 19, pp. 236–248, 2012.

[12] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viegever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[13] D. Wang, A. Haytham, J. Pottenburgh, O. Saeedi, and Y. Tao, "Hard attention net for automatic retinal vessel segmentation," *IEEE J. Biomed. Health Informat.*, to be published.

[14] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 109–118, Jan. 2016.

[15] Z. Yan, X. Yang, and K. -T. Cheng, "A three-stage deep learning model for accurate retinal vessel segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1427–1436, Jul. 2019.

[16] A. D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203–210, Mar. 2000.

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

[18] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A. K. Davison, and R. Martí, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 4, pp. 1218–1226, 2018.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[20] B. Wu *et al.*, "Generalization in generative adversarial networks: A noval perspective from privacy protection," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 307–317.

[21] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 3–18.

[22] Y. Liu, J. Peng, J. J. Q. Yu, and Y. Wu, "PPGAN: Privacy-preserving generative adversarial network," in *Proc. IEEE Int. Conf. Parallel Distrib. Syst.*, 2019, pp. 985–989.

[23] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[24] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA Cancer J. Clin.*, vol. 70, no. 1, pp. 7–30, 2020.

[25] A. Stangelberger, M. Waldert, and B. Djavan, "Prostate cancer in elderly men," *Rev. Urol.*, vol. 10, no. 2, pp. 111–119, 2008.

[26] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "Computer-aided detection of prostate cancer in MRI," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1083–1092, 2014, Feb. 2017.

[27] L. Geert, D. Oscar, B. Jelle, K. Nico, and H. Henkjan, *Prostatex Challenge Data. The Cancer Imaging Archive*, Feb. 2017. [Online]. Available: https://doi.org/10.7937/K9TCIA.2017.MURS5CL

[28] K. Clark *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, pp. 1045–1057, 2013.

[29] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[30] Z. Yan, X. Yang, and K. -T. Cheng, "A skeletal similarity metric for quality evaluation of retinal vessel segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 4, pp. 1045–1057, Apr. 2018.

[31] Z. Yan, X. Yang, and K. -T. Cheng, "Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1912–1923, Sep. 2018.