Personalized Retrogress-Resilient Federated Learning Toward Imbalanced Medical Data

Zhen Chen, *Member, IEEE*, Chen Yang, Meilu Zhu, Zhe Peng, and Yixuan Yuan, *Member, IEEE*

***Abstract*— Clinically oriented deep learning algorithms, combined with large-scale medical datasets, have significantly promoted computer-aided diagnosis. To address increasing ethical and privacy issues, Federated Learning (FL) adopts a distributed paradigm to collaboratively train models, rather than collecting samples from multiple institutions for centralized training. Despite intensive research on FL, two major challenges are still existing when applying FL in the real-world medical scenarios, including the performance degradation (*i.e.*, retrogress) after each communication and the intractable class imbalance. Thus, in this paper, we propose a novel personalized FL framework to tackle these two problems. For the retrogress problem, we first devise a Progressive Fourier Aggregation (PFA) at the server side to gradually integrate parameters of client models in the frequency domain. Then, at the client side, we design a Deputy-Enhanced Transfer (DET) to smoothly transfer global knowledge to the personalized local model. For the class imbalance problem, we propose the Conjoint Prototype-Aligned (CPA) loss to facilitate the balanced optimization of the FL framework. Considering the inaccessibility of private local data to other participants in FL, the CPA loss calculates the global conjoint objective based on global imbalance, and then adjusts the client-side local training through the prototype-aligned refinement to eliminate the imbalance gap with such a balanced goal. Extensive experiments are performed on real-world dermoscopic and prostate MRI FL datasets. The experimental results demonstrate the advantages of our FL framework in real-world medical scenarios, by outperforming state-of-the-art FL methods with a large margin. The source code is available at https://github.com/CityU-AIM-Group/PRR-Imbalancehttps://github.com/CityU-AIM-Group/PRR-Imbalance.**

***Index Terms*— Federated learning, retrogress, class imbalance, dermoscopic diagnosis, prostate segmentation.**

I. Introduction

In the field of computer-aided diagnosis, advanced deep learning techniques have provided clinicians with reliable and effective assistance for patient diagnosis [1], [2]. The great success of these works relies on the valuable knowledge of large-scale datasets collected from multiple institutions. However, constructing increasingly-large centralized datasets is not sustainable for future smart healthcare systems, owing to patient privacy and ethical concerns. To protect data privacy, Federated Learning (FL) [3], [4], as a distributed machine learning framework, enables multiple data owners to conduct local model training and aggregate models on a central server in an iterative fashion. Without revealing the concrete data, the model aggregated with global knowledge can outperform the model optimized on any single client. Among existing FL works, personalized FL [5]–[7] is particularly suitable for medical scenarios, since each client is allowed to choose either the aggregated server model or local models according to its preference. However, two challenges will be met, when FL is applied in the real-world medical scenarios.

For real-world medical FL, differences in imaging devices, protocols and regional diseases are likely to cause severe data heterogeneity among clients. The resulting enormous discrepancy between different local models might disable certain operations in existing FL works. With F1 training curves visualized in Fig. 1 (a), we notice that classic FL works (*e.g.*, FedAvg [8]) encounter a violent descent of performance after each server-client communication, which is termed as *retrogress* [9]. The aggregated models with retrogress obliterate the previous knowledge and need to be re-adapted to the client task in the next local training. This would degrade the local training of clients and the knowledge sharing of the server. Moreover, the deputy model in the personalized FML [5] also suffers from the retrogress problem, and further degrades the performance of the local model, which results in an inferior curve compared to ours in Fig. 1 (a). Specifically, we suppose the retrogress phenomenon may come from the following two reasons. On the one hand, due to the severe inter-client data heterogeneity, parameters of different client models may represent diverse semantic patterns at the same position [10]. It is not reasonable to average these parameters in element-wise. In contrast, converting parameters to frequency domain can effectively align the parameter components along the frequency dimension, and provides the feasibility to choose the

Manuscript received 31 March 2022; revised 28 May 2022; accepted 13 July 2022. Date of publication 19 July 2022; date of current version 2 December 2022. This work was supported in part by the Shenzhen-Hong Kong Innovation Circle Category D Project under Grant SGDX2019081623300177 (CityU 9240008) and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020A1515111070. *(Corresponding author: Yixuan Yuan.)*

Zhen Chen, Chen Yang, Meilu Zhu, and Yixuan Yuan are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong, SAR, China (e-mail: zchen.ee@my.cityu.edu.hk; cyang.ee@my.cityu.edu.hk; meiluzhu2-c@my.cityu.edu.hk; yxyuan.ee@cityu.edu.hk).

Zhe Peng is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China (e-mail: pengzhe@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/TMI.2022.3192483

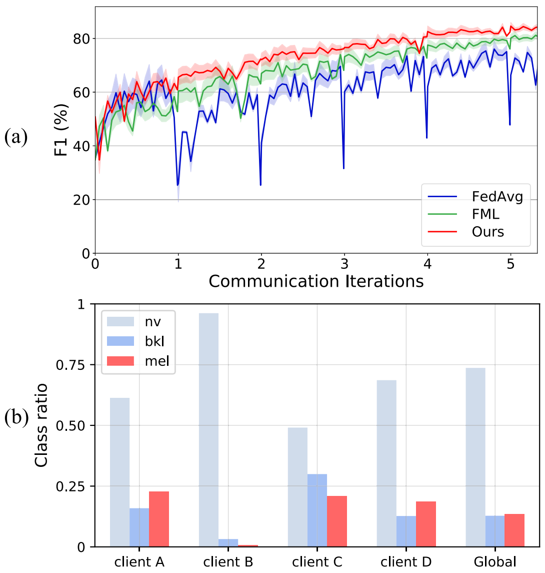

Fig. 1. (a) The F1 training curves of local models at client C in dermoscopic FL dataset. Our method effectively alleviates the retrogress after each client-server communication, leading to superior FL performance. (b) The class imbalance of real-world dermoscopic FL dataset. The nevus (*nv*), benign keratosis (*bkl*) and melanoma (*mel*) are severely unbalanced, and the significant distribution gap exists between clients.

aggregated frequency band [11]. On the other hand, replacing the previous local model with the aggregated server model would discard the previous knowledge learned by local models and hinder the optimization in the next iteration.

Another challenge is that the *class imbalance* in medical imaging becomes more complicated in the FL scenarios, which may cause great impacts on local training and server aggregation. The class ratio of the real-world dermoscopic FL dataset in Fig. 1 (b) reveals the obvious imbalance between categories and clients. First, medical imaging is inherently unbalanced due to the scarcity of target diseases, *e.g.*, keratosis and melanoma are relatively rarer than nevus among skin lesion samples. Second, there exist significant differences in the category distribution among clients in real-world FL. Compared with the European population, the incidence of melanoma among Australian patients is relatively higher due to severe chronic sunburn [12]. Furthermore, the inaccessibility of decentralized training samples to other participants makes the class imbalance in FL more difficult to deal with. To tackle the class imbalance problem in FL, a rational idea is to first formulate a balanced goal based on the global imbalance of the FL framework and then adjust the local training of each client to eliminate the imbalance gap between the client model and the balanced objective in a collaborative manner.

To address the aforementioned retrogress and class imbalance problems, we propose a personalized retrogress-resilient FL framework to generate a superior personalized model for each client. Specifically, we first propose a Progressive Fourier Aggregation (PFA) to integrate client models at the server side. With the mutual conversion of client parameters to frequency domain by Fast Fourier Transform (FFT) and inverse FFT (IFFT), we average the low-frequency components of client parameters while retaining the individual high-frequency components. By increasing the frequency threshold of shared components gradually during FL training, PFA can effectively integrate client knowledge in a manner consistent with network learning preferences. Then, instead of replacing local models at the client side, we devise a Deputy-Enhanced Transfer (DET) to introduce a deputy model to receive the updated server model and maintain the personalized local model without contamination. To address the retrogress of deputy model, the DET restores the local prior for the deputy model and transfers the global knowledge to promote the personalized local model through the tailored Recover-Exchange-Sublimate steps. Moreover, we devise the Conjoint Prototype-Aligned (CPA) loss to overcome the class imbalance problem in FL. From the server perspective, the global conjoint objective is calculated based on the global imbalance of the FL framework. Towards such a balanced goal, the local prototype-aligned refinement adjusts the local training in each client to eliminate the imbalance gap, by restricting the difference between the local prototype and the global prototype from the server.

The contributions of this work are summarized as follows:

- To handle real-world medical FL, we propose a novel personalized retrogress-resilient FL framework with improved Progressive Fourier Aggregation PFA) at the server side. Particularly, we design an improved strategy for frequency component aggregation and the category-wise operation for the classifier's parameters.
- We devise the Deputy-Enhanced Transfer (DET) to improve the personalized local model with Recover-Exchange-Sublimate steps at the client side, which can transfer the global knowledge without being interrupted by the server-client communication.
- Aiming at the class imbalance problem in FL scenarios, we propose the Conjoint Prototype-Aligned (CPA) loss from the two aspects of global conjoint objective and local prototype-aligned refinement. As such, the FL framework can be optimized towards a balanced goal.
- Plenty of FL experiments on dermoscopic diagnosis and prostate MRI segmentation prove the effectiveness of our personalized retrogress-resilient FL framework, outperforming state-of-the-art FL works by a large margin.

A preliminary version of this work has been published in MICCAI 2021 [9]. In this paper, we have made a significant extension with the following highlights: 1) We devise the Conjoint Prototype-Aligned (CPA) loss to handle the imbalance problem in real-world medical FL; 2) Compared with the PFA in the conference work [9], we further aggregate the phase components to integrate sufficient knowledge, and independently process the category-wise parameters of the classifier to avoid potential bias; 3) Besides classification experiments on dermoscopic images, we conduct extensive experiments to enhance the comprehensive validation, including segmentation experiments on prostate MRI dataset, comparison with existing FL imbalance works and detailed ablation studies.

II. Related Work

A. Federated Learning

Federated Learning (FL) is a decentralized framework for collaborative optimization of client models while preserving private data. Particularly, FedAvg [8] first proved that averaging local models of different clients is equivalent to updating the gradients in a centralized manner, when clients are assumed as independent and identically distributed (IID). To alleviate the practical non-IID distribution between clients, FedProx [13] constrained the heterogeneous local updates close to the global model in the previous iteration, by introducing a parameter regularization term to the local training objective. To overcome the inter-client data heterogeneity, FedBN [14] and SiloBN [7] employed personalized normalization statistics for client models, and FML [5] introduced the deputy model at the client side to perform mutual learning [15] between the deputy model and the personalized model. More recently, FedProto [16] utilized category prototypes instead of local gradients to perform the server-client communication, and constrained the local prototypes to be sufficiently close to the global prototype aggregated on the server.

In computer-aided diagnosis, FL with increasing attention has been applied to a variety of medical tasks for exploratory attempts [17]. Li *et al.* [18] implemented a practical FedAvg [8] system with differential-privacy techniques on brain MRI segmentation. Dou *et al.* [19] investigated the feasibility of detecting Coronavirus Disease 2019 (COVID-19) abnormalities from lung CT scans among multi-national hospitals. To utilize the knowledge of unlabeled data at the client side, Yang *et al.* [20] extended the FL framework to semi-supervised CT segmentation by imposing consistency between pseudo labels and the predictions after input perturbations. IDA [21] adopted the parameter changes as the client importance for server aggregation, and this weighting strategy was validated on the dermoscopic FL dataset with randomly sampled clients. Different from previous FL works in medical imaging, we systematically analyze the challenges of real-world medical FL tasks, and propose the tailored solution to handle two main problems, including the retrogress and class imbalance.

B. Class Imbalance

Recently, class imbalance has attracted a lot of attention as a significant issue for real-world applications, including classification [22], segmentation [23] and detection [24]. In the imbalanced dataset, the dominant categories with more samples would overwhelm the rare categories, which severely degrades the performance of biased networks. Most imbalance research can be grouped into re-sampling [25]–[27] and re-weighting [22]–[24]. According to the degree of class imbalance, re-sampling works aim to adjust the category distribution towards a re-balanced goal, *e.g.*, over-sampling for rare categories, under-sampling for dominant categories and class-balanced sampling [25]. By treating the network as two parts, the recent decoupling strategy [26], [27] was adopted to first optimize the feature extractor with instance-balanced sampling and then re-train the classifier with class-balanced sampling. These re-sampling approaches usually employ different data samplers to formulate a multi-stage training procedure, which is relatively cumbersome for the FL scenarios.

To avoid the network being biased to the dominant categories, re-weighting works introduced larger penalties on the rare or error-prone categories. Particularly, focal loss [24] was proposed to adaptively increase the penalty of hard samples while overlooking the easy negatives. Recent works [22], [23] pointed out that the competition gradients produced by dominant categories severely inhibit the network learning of rare categories, and significantly improved the performance by excluding such competition gradients of rare category samples. For the medical imaging, the loss re-weighting and sample re-sampling based on the inverse of class frequency were widely used to alleviate the imbalance problem [28]. Compared with existing class imbalance works focusing on centralized scenarios, our method formulates the more complicated imbalance problem in FL scenarios into the global imbalance and the local imbalance gap, and further handles this problem under the protection of private data at the client side.

III. Methodology

A. Overview

Given K clients with private data, the proposed personalized retrogress-resilient FL framework aims to collaboratively produce a personalized local model with superior performance for each client. These models $\{\boldsymbol{P}_k\}_{k=1}^K$ share the same network architecture to benefit from the server aggregation. For the k-th client, the personalized local model $\boldsymbol{P}_k$ is optimized with private data for E-epoch local training. Then, the server collects these trained client models, and utilizes the Progressive Fourier Aggregation (PFA) to aggregate them into server models $\{\boldsymbol{S}_k\}_{k=1}^K$ with individual high-frequency components. After that, these server models are delivered to the corresponding client as a deputy model $\boldsymbol{D}_k$. The deputy model can transfer the global knowledge through the proposed Deputy-Enhanced Transfer (DET). In addition, by utilizing local prototypes at client and global prototype at server, the Conjoint Prototype-Aligned (CPA) loss guides the local training of clients towards a global balanced goal to tackle the class imbalance problem in FL. Repeat these steps until local training reaches T epochs. We demonstrate the personalized retrogress-resilient FL framework in Fig. 2.

B. Progressive Fourier Aggregation in Server

To integrate the global knowledge from client models, existing FL methods [7], [8], [13], [14] directly averaged the parameters of local models in element-wise to generate the aggregated server model. However, this rough aggregation at parameter space abruptly degrades the model performance on clients, as the retrogress of FedAvg [8] shown in Fig. 1. To address the retrogress caused by aggregation, we devise the Progressive Fourier Aggregation (PFA) to stably integrate the global knowledge at the frequency domain.

Considering that low-frequency components of parameters are the basis for network capability [11], our PFA aggregates the relatively low-frequency components of parameters to share knowledge from different clients, while preserving their high-frequency components containing specific knowledge for each individual client. Specifically, for a convolutional layer in k-th client model, we first reshape its 4-D parameter tensor into a 2-D matrix $w_k \in \mathbb{R}^{h_1 d_1 \times h_2 d_2}$, where d_1 and d_2 are

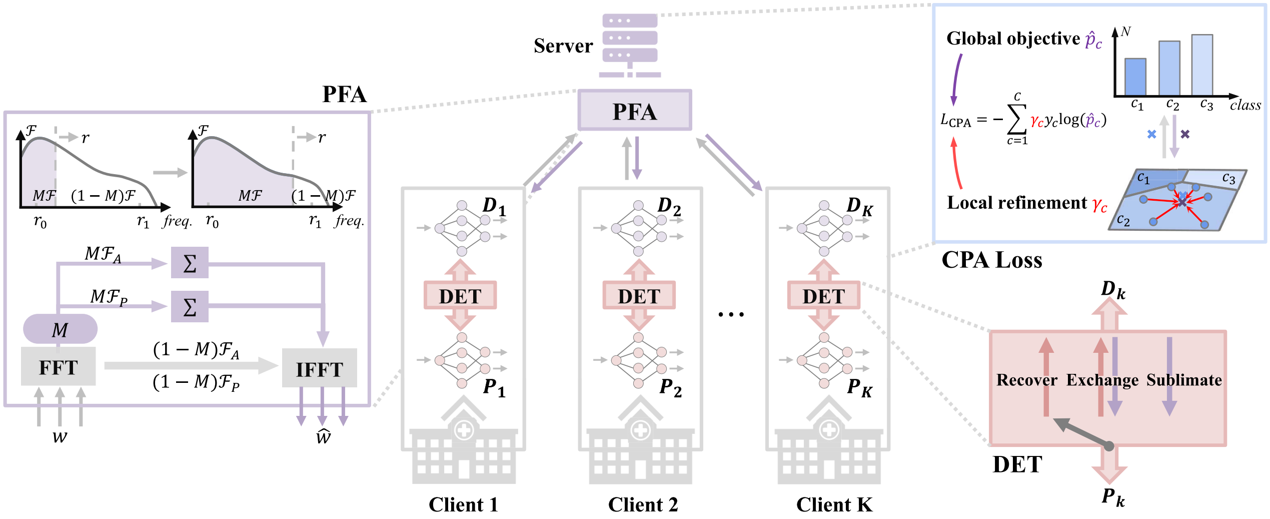

Fig. 2. The personalized retrogress-resilient FL framework. At the server side, Progressive Fourier Aggregation (PFA) integrates global knowledge gradually in the frequency domain. At the client side, Deputy-Enhanced Transfer (DET) promotes the personalized local model without being interrupted by communications. The Conjoint Prototype-Aligned (CPA) loss guides local training of each client towards a global balanced goal.

the output and input channels, and h_1 and h_2 are the kernel spatial size. Then, the amplitude components $\mathcal{F}_A$ and phase components $\mathcal{F}_P$ of parameters are generated by the Fourier transform $\mathcal{F}(w_k) = \mathcal{F}_A(w_k)e^{\mathbf{j}\mathcal{F}_P(w_k)}$, as follows:

$$\mathcal{F}(w_k)(m,n) = \sum_{x,y} w_k(x,y)e^{-\mathbf{j}2\pi\left(\frac{x}{h_1 d_1}m + \frac{y}{h_2 d_2}n\right)}, \quad (1)$$

where $\mathbf{j}^2 = -1$. This procedure can be efficiently conducted with FFT [29]. To extract the low-frequency components for aggregation, we adopt a low-frequency mask M with zero-value except for the central area:

$$M(m,n) = \mathbb{1}_{(m,n)\in[-rh_1d_1:rh_1d_1,\,-rh_2d_2:rh_2d_2]}, \quad (2)$$

where $r \in (0, 0.5)$ is the low-frequency threshold, and the center of w_k is set as the coordinate $(0, 0)$. To gather global knowledge from client models, the amplitude components and phase components of k-th client are aggregated by averaging the low-frequency components over K clients, respectively:

$$\hat{\mathcal{F}}_A(w_k) = (1-M)\circ\mathcal{F}_A(w_k) + \frac{1}{K}\sum_{i=1}^{K} M\circ\mathcal{F}_A(w_i), \quad (3)$$

$$\hat{\mathcal{F}}_P(w_k) = (1-M)\circ\mathcal{F}_P(w_k) + \frac{1}{K}\sum_{i=1}^{K} M\circ\mathcal{F}_P(w_i), \quad (4)$$

where $\circ$ is the element-wise multiplication. Motivated by the fact that networks tend to learn the low-frequency knowledge prior to the high-frequency counterpart [30], we implement the PFA with a progressive strategy by increasing $r = r_0 + \frac{r_1 - r_0}{T}t$ in Eq. (2) during the FL training, where r_0 and r_1 are the initial and terminated low-frequency threshold. By converting the amplitude and phase components back to parameters using the inverse Fourier transform $\mathcal{F}^{-1}$, we finally obtain the aggregated parameters of k-th client as $\hat{w}_k = \mathcal{F}^{-1}([\hat{\mathcal{F}}_A(w_k), \hat{\mathcal{F}}_P(w_k)])$.

Besides the aforementioned analysis for convolutional layers, PFA can be easily applied to the fully-connected (FC) layers of client models without reshaping the 2-D parameters. A special case is that for the last FC layer, once the training data is imbalanced, the category-wise parameters hold different magnitudes of norms [26]. Therefore, directly performing PFA on the parameters of the last FC layer may ignore the categories with fewer samples. Instead, we implement PFA on the category-wise parameters of the last FC layer independently, by performing FFT/IFFT and mask operation on one dimension.

C. Deputy-Enhanced Transfer in Client

The PFA can overcome the retrogress caused by improper server aggregation, however, directly replacing client models with the aggregated server parameters still loses the previous local knowledge and further interferes with the optimization in the next iteration. To overcome this bottleneck, we devise the Deputy-Enhanced Transfer (DET) to merge global knowledge with local priors, instead of the direct replacement. Besides the personalized local model $\boldsymbol{P}$, each client holds a deputy model $\boldsymbol{D}$ to receive the aggregated parameters from the server. By conducting three steps of Recover-Exchange-Sublimate, the proposed DET smoothly transfers global knowledge from the deputy model $\boldsymbol{D}$ to the personalized local model $\boldsymbol{P}$.

1) Recover: When updated with the aggregated model $\boldsymbol{S}$ from the server, the deputy model $\boldsymbol{D}$ is severely degraded by the retrogress. Therefore, we firstly adopt the personalized local model $\boldsymbol{P}$ as a teacher to restore the deputy model $\boldsymbol{D}$ with the local knowledge. In this step, the personalized local model $\boldsymbol{P}$ is optimized with L_{sup} for the independent local training, while the deputy model $\boldsymbol{D}$ is optimized by the following loss function $L_{\boldsymbol{D}}$:

$$L_{\boldsymbol{D}} = L_{sup} + L_{KL}(p_{\boldsymbol{P}} \| p_{\boldsymbol{D}}), \quad (5)$$

where L_{sup} is the supervision loss of local dataset, *e.g.*, cross entropy loss or the proposed CPA loss, and $p_{\boldsymbol{P}}$ and $p_{\boldsymbol{D}}$ are the predicted probabilities of the personalized local model $\boldsymbol{P}$ and the deputy model $\boldsymbol{D}$. The $L_{KL}(p_{\boldsymbol{P}} \| p_{\boldsymbol{D}})$ is the Kullback-Leibler divergence to help the deputy model quickly re-adapt to the client with performance improved. This step is crucial to ensure the deputy model does not harm the personalized local model in the subsequent transfer of global knowledge.

2) Exchange: Once the recovered performance of the deputy model $\boldsymbol{D}$ is close to the teacher $\boldsymbol{P}$, as $\phi_{val}(\boldsymbol{D}) \geq \lambda_1 \phi_{val}(\boldsymbol{P})$, where ϕ_{val} represents a specific performance metric on the validation set (*e.g.*, F1 for classification and Dice for segmentation), we perform the mutual learning [15] between the personalized local model $\boldsymbol{P}$ and the deputy model $\boldsymbol{D}$ to exchange the global knowledge and the local knowledge. Specifically, the deputy model $\boldsymbol{D}$ is optimized by Eq. (5) and the personalized local model $\boldsymbol{P}$ is supervised by the loss $L_{\boldsymbol{P}}$:

$$L_{\boldsymbol{P}} = L_{sup} + L_{KL}\left(p_{\boldsymbol{D}} \| p_{\boldsymbol{P}}\right). \tag{6}$$

The $L_{KL}\left(p_{\boldsymbol{D}} \| p_{\boldsymbol{P}}\right)$ term in Eq. (6) smoothly transfers the global knowledge of the server to the personalized local model $\boldsymbol{P}$ through the deputy model $\boldsymbol{D}$. Therefore, the knowledge exchange can improve the generalization of client models [15].

3) Sublimate: Finally, when the performance of the deputy model is highly close to the personalized local model, as $\phi_{val}(\boldsymbol{D}) \geq \lambda_2 \phi_{val}(\boldsymbol{P})$, where $0 < \lambda_1 < \lambda_2 < 1$, the deputy model $\boldsymbol{D}$ maintains the independent local training with L_{sup}, and further serves as the teacher to optimize $\boldsymbol{P}$ with $L_{\boldsymbol{P}}$ of Eq. (6). This step allows the global knowledge can be transferred from the server to the personalized local model to the greatest extent.

D. Conjoint Prototype-Aligned Loss for FL Imbalance

The class imbalance severely degrades the performance of deep learning methods [22], [23]. In addition to the inherent imbalance in medical imaging, the real-world FL also involves significant differences in category distribution among clients. Particularly, this issue is difficult to deal with when decentralized training samples are inaccessible to other participants. To tackle the complicated class imbalance problem in FL, we devise the Conjoint Prototype-Aligned (CPA) loss by guiding the local training of clients towards a global balanced goal, as shown in Fig. 3. Specifically, we first calculate the global conjoint objective based on the imbalance of the entire FL framework, and then adjust the local training of each client to eliminate the prototype-measured imbalance gap between local models and the balanced objective.

1) Global Conjoint Objective: We first revisit the widely used cross entropy loss. Denote $\boldsymbol{z} = [z_1, z_2, \ldots, z_C]$ as the output logits for C-category, the cross entropy loss $L_{\mathrm{CE}}(\boldsymbol{z})$ is calculated as follows:

$$L_{\mathrm{CE}}(\boldsymbol{z}) = -\sum_{c=1}^{C} y_c \log\left(p_c\right), \tag{7}$$

where $y_c \in \{0, 1\}$, $1 \leq c \leq C$ is the one-hot ground-truth label, and $p_c = \exp(z_c)/\sum_{j=1}^{C} \exp(z_j)$ represents the predicted probability of the c-th category adjusted by the softmax function. Given a c-th category sample, the gradients on z_c, $\frac{\partial L_{\mathrm{CE}}(\boldsymbol{z})}{\partial z_c} = p_c - 1$, optimize the network parameters to correctly classify this sample. Although the j-th category prediction z_j does not contribute to the intensity of loss L_{CE} in Eq. (7), the gradients on z_j will suppress the prediction of z_j due to the softmax competition among logits, as follows:

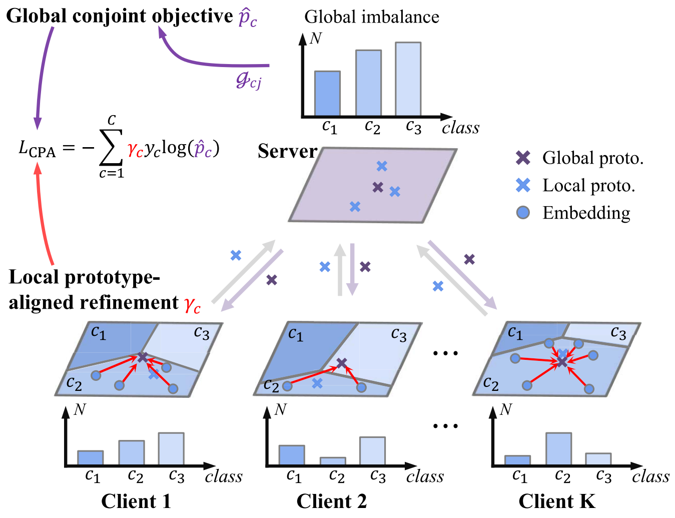
Global conjoint objective $\hat{p}_c$
Global imbalance
N
c_1 c_2 c_3 class
$\mathcal{G}_{cj}$
$L_{\mathrm{CPA}} = -\sum_{c=1}^{C} \gamma_c y_c \log(\hat{p}_c)$
Server
Global proto.
Local proto.
Embedding
Local prototype-aligned refinement γ_c
c_1 c_2 c_3
...
Client 1
Client 2
Client K

Fig. 3. The Conjoint Prototype-Aligned (CPA) loss for FL imbalance. At server, the global conjoint objective proposes a balanced goal for the entire FL framework based on the global imbalance. At client, the local prototype-aligned refinement eliminates the imbalance gap during local training by restricting local prototypes (*i.e.*, feature embeddings as Eq. (10) for better visualization) towards corresponding global prototype.

$$\frac{\partial L_{\mathrm{CE}}(\boldsymbol{z})}{\partial z_j} = p_j. \tag{8}$$

In the imbalanced case, the rare categories will be heavily suppressed by enormous samples of dominant categories, as the competition gradients in Eq. (8). This forces the network to be biased towards the dominant categories, thereby leading to poor performance, especially on rare categories [22], [23].

For the global conjoint objective, we modify the softmax function with the mask $\mathcal{G}_{cj} = \min(1, \left(N_j/N_c\right)^{\beta})$ to reduce the competition gradients on rare categories, with the modified probability $\widehat{p}_c$ as follows:

$$\widehat{p}_c = \frac{\exp(z_c)}{\exp(z_c) + \sum_{j \neq c}^{C} \mathcal{G}_{cj} \exp(z_j)}, \tag{9}$$

where N_c and N_j are the training sample numbers of the c-th and j-th categories in the entire FL framework, and β is the coefficient to balance the adjustment among categories. When $N_j < N_c$, the rare j-th category receives the reduced suppression from the dominant c-th category based on sample counts, with the competition gradients $\widehat{p}_c \frac{\exp(z_j)}{\exp(z_c)} (\frac{N_j}{N_c})^{\beta}$ adjusted by the mask $\mathcal{G}_{cj} = (\frac{N_j}{N_c})^{\beta}$. Once $N_j \geq N_c$, the incorrect j-th category with more samples receives the consistent competition from the rare c-th category as vanilla softmax function, which ensures accurate prediction. Note that the sample numbers N_c and N_j are collected from clients and demand negligible communication overhead [16], [31]. In this way, the global conjoint objective in Eq. (9) serves as a balanced goal for local training to correct the FL imbalance from a global perspective.

2) Local Prototype-Aligned Refinement: Towards the balanced goal in Eq. (9), we aim to eliminate the imbalance gap between the local training in each client and the global

Algorithm 1: The Pipeline of Our FL Framework

```
Input  : The FL dataset and category distribution;
         The network structure for $\boldsymbol{P}$, $\boldsymbol{D}$, $\boldsymbol{S}$;
         Local epochs $E$;
         Total epochs $T$;
Output: The trained personalized local models $\{\boldsymbol{P}_k\}_{k=1}^{K}$.
Server executes:
  Initialize the model parameters $\boldsymbol{P}$, $\boldsymbol{D}$, $\boldsymbol{S}$ for all clients
    and global prototype $\bar{\boldsymbol{f}}$ for all categories;
  Calculate $\mathcal{G}_{cj}$ based on category distribution for global
    conjoint objective in Eq. (9);
  for each round in $T/E$ do
    for each client $k$ in parallel do
      $\boldsymbol{D}_k$, $\boldsymbol{f}^{(k)}$ $\leftarrow$ LocalUpdate($k$, $\boldsymbol{S}_k$, $\bar{\boldsymbol{f}}$);
    end for
    Generate the aggregated model $\boldsymbol{S}_k$ with PFA;
    Update global prototype $\bar{\boldsymbol{f}}$;
  end for
LocalUpdate($k$, $\boldsymbol{S}_k$, $\bar{\boldsymbol{f}}$):
  Update $\boldsymbol{D}_k$ with $\boldsymbol{S}_k$;
  for each local epoch in $E$ do
    Compute local prototype $\boldsymbol{f}^{(k)}$ in Eq. (10);
    Compute $L_{\mathrm{CPA}}$ of $\boldsymbol{D}_k$ and $\boldsymbol{P}_k$ in Eq. (13);
    Optimize $\boldsymbol{D}_k$ and $\boldsymbol{P}_k$ with DET using $L_{\mathrm{CPA}}$.
  end for
  return $\boldsymbol{D}_k$ and $\boldsymbol{f}^{(k)}$
```

conjoint objective. Motivated by the category-wise representation differences caused by imbalanced datasets [31], we utilize the prototype differences to measure the imbalance gap, and introduce the prototype-aligned refinement to adjust the category-wise penalty of loss function in local training. Specifically, we first collect the category-wise prototype of each client. For the k-th client, the local prototype $\boldsymbol{f}_c^{(k)}$ is calculated by averaging the embeddings of $N_c^{(k)}$ samples belonging to c-th category:

$$\boldsymbol{f}_c^{(k)} = \frac{1}{N_c^{(k)}} \sum_{i=1}^{N_c^{(k)}} f(x_i), \tag{10}$$

where $f(x_i)$ is the feature embedding of c-th category sample, and we empirically adopt the feature vector before the last fully-connected layer [16]. With the c-th category local prototypes $\{\boldsymbol{f}_c^{(k)}\}_{k=1}^{K}$ of clients delivered to the server, the global prototype $\bar{\boldsymbol{f}}_c$ is sampled from the Gaussian distribution $\bar{\boldsymbol{f}}_c \sim \mathcal{N}(\mu, \sigma^2)$, where mean value $\mu = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{f}_c^{(k)}$ and variance $\sigma^2 = \frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{f}_c^{(k)} - \mu)^2$. Different from previous works [16], [31] that directly averaged client prototypes as the global prototype, our method leverages the sampling augmentation with second-order statistical information of client prototypes. Then, by adopting the cosine similarity between $\boldsymbol{f}_c^{(k)}$ and $\bar{\boldsymbol{f}}_c$ to measure the imbalance gap, the prototype distance $\gamma_c^{(k)}$ is calculated to refine the category-wise penalty:

$$\gamma_c^{(k)} = \Gamma\left(\frac{\boldsymbol{f}_c^{(k)} \cdot \bar{\boldsymbol{f}}_c}{\|\boldsymbol{f}_c^{(k)}\|_2 \|\bar{\boldsymbol{f}}_c\|_2}, \tau\right), \tag{11}$$

where $\cdot$ represents the inner product of vectors, and Γ is the distance function to produce a larger term when the similarity is lower, as follows:

$$\Gamma(s, \tau) = \frac{1+\tau}{s+\tau}, \tag{12}$$

where τ is a hyper-parameter to control the change of Γ w.r.t the variable s. When the local prototype $\boldsymbol{f}_c^{(k)}$ and the global prototype $\bar{\boldsymbol{f}}_c$ are not aligned, $\gamma_c^{(k)} > 1$ leads to more penalties on this category until the imbalance gap is eliminated.

3) CPA Loss Calculation: To mitigate the imbalance gap between each client and the global conjoint objective, we calculate the CPA loss L_{CPA} by emphasizing these categories with larger differences in local and global prototypes:

$$L_{\mathrm{CPA}}(z) = -\sum_{c=1}^{C} \gamma_c y_c \log\left(\widehat{p}_c\right), \tag{13}$$

where $\widehat{p}_c$ is the refined probability in Eq. (9). When prototype differences exist between the local client and the global conjoint objective, $\gamma_c > 1$ and L_{CPA} will impose more penalties on this category until the local prototype is aligned to corresponding global one. Ideally, when local and global prototypes are aligned, γ_c is equal to 1, which means the consistent category-wise representations between the local model and the global model. By optimizing the L_{CPA} loss in Eq. (13), the local models in each client are adaptively improved towards the balanced goal, thereby overcoming the imbalance problem of the entire FL framework in a collaborative manner.

E. Optimization Pipeline

We summarize our FL framework in Algorithm 1. In each iteration, the server receives the deputy model $\boldsymbol{D}$ and local category prototype $\boldsymbol{f}$ from each client, and then aggregates these client models with PFA in Eq. (3) and Eq. (4) and updates the global prototype $\bar{\boldsymbol{f}}$. As the aggregated models are sent back to the corresponding client, the overridden deputy model $\boldsymbol{D}$ first recovers its performance on the private local data, and then transfers the global knowledge smoothly to the personalized local model $\boldsymbol{P}$ with the DET. In particular, the CPA loss L_{CPA} in Eq. (13) serves as the supervision loss L_{sup} in Eq. (5) and Eq. (6) during the DET local training, and eliminates the imbalance of FL task by emphasizing those categories with large prototype differences. Finally, the personalized local models $\{\boldsymbol{P}_k\}_{k=1}^{K}$ are well optimized towards the balanced goal.

IV. Experiment

A. Datasets and Implementations

1) Real-World Dermoscopic FL Dataset: To evaluate the performance of FL frameworks in real-world medical scenarios with challenging heterogeneity, we build a FL benchmark dataset for dermoscopic diagnosis based on the practical client setting. From HAM10K [12] dataset and MSK [32] dataset respectively, we collect 8, 940 and 2, 000 samples belonging to nevus (*nv*), benign keratosis (*bkl*) or melanoma (*mel*). To retain the critical lesion shape information, we first crop the central square areas with shorter sides from the irregularly sized images, and then resize the resolution to 128 × 128. According

TABLE I
REAL-WORLD FL DATASET INFORMATION

Client	A	B	C	D	OOD
Cohort	ViDIR Modern	ViDIR MoleMax	Rosendahl	MSK	ViDIR Legacy
Case num	2,987	3,868	1,635	2,000	427
nv	1,832	3,720	803	1,372	350
bkl	475	124	490	254	10
mel	680	24	342	374	67
Prostate MRI FL Dataset					
Client	A	B	C	D	E
Cohort	RUNMC	BMC	UCL	BIDMC	HK
Device	Siemens	Philips	Siemens	GE	Siemens
Case num	30	30	13	12	12
V_P (%)	2.65	1.76	1.45	2.16	1.39
V_B (%)	97.35	98.24	98.55	97.84	98.61

to image metadata [12], we partition samples from the same cohort into one client. As shown in Table I, the dermoscopic FL dataset includes four clients, where the first three clients are from HAM10K [12] and MSK [32] data serves as one client. We further split the samples of each client into training set, validation set and test set as 7:1:2, and guarantee the same category distribution among three sets. Moreover, an unseen cohort with 427 samples is also collected to evaluate the out-of-distribution (OOD) generalization in Section IV-E.

2) Real-World Prostate MRI FL Dataset: To further perform evaluation on the segmentation task, we also collect a real-world prostate MRI FL dataset from NCI-ISBI 2013 [33] and PROMISE12 [34] datasets. With institution information, these T2-weighted MRI samples are segregated into five clients, with client A and B from NCI-ISBI 2013 [33] and client C, D and E from PROMISE12 [34]. As illustrated in Table I, we calculate voxel ratios of the prostate (V_P) and background area (V_B) for each client, which demonstrates the severe imbalance in the segmentation task and the imbalance gap among five clients. Following the pre-processing in [35], we unify the axial plane slices into resolution of 256×256. The samples of each client are divided into training, validation and test sets as 7:1:2 at the patient level. In this way, these two datasets involve data and label heterogeneity, and reveal the challenges of real-world medical FL, such as various imaging devices, different sample numbers of clients and class imbalance in FL.

3) Implementations: We implement our FL framework and state-of-the-art FL approaches on NVIDIA V100 GPU using PyTorch [36], and adopt VGG-16BN [37] for dermoscopic diagnosis and Deeplab-v2 [38] with ResNet-18 [39] backbone for prostate segmentation. For the local training of dermoscopic FL, the networks are optimized with SGD for $T = 250$ epochs, with initial learning rate 1×10^{-2} and batch size 16. For the prostate segmentation task, we optimize the networks using Adam for $T = 60$ epochs, with initial learning rate 1×10^{-3} and batch size 8. On both datasets, the server-client communication is performed after every $E = 5$ epochs during the local training, and the learning rate is divided by 2 after every 25 epochs. During the communication, the deputy model in our framework retains the personalized BN layers [14]. The frequency thresholds r_0 and r_1 in PFA are set as 0.35 and 0.48, respectively. In DET, we check for step switching after each epoch of local training, and the F1 and Dice are chosen as the metric ϕ_{val} for classification and segmentation tasks to measure the performance of client models, with performance thresholds $\lambda_1 = 0.7$ and $\lambda_2 = 0.9$. In the CPA loss, we set hyper-parameters $\beta = 0.8$ and $\tau = 3$. In the segmentation experiment, we further employ dice loss [40] to provide regional supervision for all FL methods, together with the pixel-wise classification loss (*i.e.*, cross entropy loss or CPA loss) using the loss weight as 1 : 1. The local training is augmented with random flip and rotation of input images to avoid overfitting. The reported model is chosen by the validation set of each client.

4) Evaluation Metrics: To comprehensively evaluate the performance of various FL methods, the F1 and AUC are utilized for the dermoscopic diagnosis task. For the three-category diagnosis, we first compute the metric for each class individually, and then calculate the macro-average to consider each category equally [41]. The higher F1 and AUC scores reveal superior diagnostic performance on the dermoscopic FL dataset. For the prostate segmentation task, we adopt Dice coefficient and intersection over union (IoU) of prostate regions to evaluate model performance. The higher Dice and IoU validate better predictions of prostate segmentation task. Besides the FL performance of each client, we further demonstrate the macro-average of different clients for a more concise comparison.

B. Comparison With State-of-the-Art FL Works

1) Dermoscopic Diagnosis: On the dermoscopic FL dataset, we compare our framework with the state-of-the-art FL works from diverse perspectives. Besides the classic FedAvg [8], we select SiloBN [7] and IDA [21] designed for medical FL, and FedProx [13], FedBN [14], FedProto [16] and FML [5] to address data heterogeneity. As illustrated in Table II, FedAvg [8] obtains average F1 of 51.64% and average AUC of 77.89%, serving as the baseline for dermoscopic FL. By aligning the category prototypes between the server and client sides, the recent FedProto [16] improves the performance on heterogeneous FL data with average F1 of 62.95% and average AUC of 84.43%, which confirms the advantage of prototypes for FL applications. In contrast, our method achieves the overwhelming performance on all clients, with the best average F1 of 72.82% and average AUC of 89.20%. Particularly, our method outperforms the personalized FL approach, FedBN [14], with a remarkable increase of 11.86% in average F1 and 4.02% in average AUC. This advantage validates that our FL framework can effectively address the retrogress in FedBN [14]. Compared with FML [5] that also introduces deputy models to perform mutual learning [15], our FL framework achieves the improved performance by a large margin, *e.g.*, 5.22% in average F1 and 1.27% in average AUC. We further illustrate the standard deviation of five independent repeated experiments in Fig. 5 (a), and our FL method reveals statistically significant performance advantages with the P-value < 0.001 over all baselines. The comparison on the dermoscopic diagnosis task confirms the performance advantage of our personalized retrogress-resilient FL framework over state-of-the-art works in real-world medical FL.

TABLE II
COMPARISON WITH STATE-OF-THE-ART FL WORKS ON REAL-WORLD DERMOSCOPIC FL DATASET

Method	F1 (%)					AUC (%)				
	A	B	C	D	Avg	A	B	C	D	Avg
FedAvg [8]	57.44	48.14	56.80	44.20	51.64	81.52	82.75	76.14	71.14	77.89
FedProx [13]	56.70	39.09	54.70	45.95	49.11	81.70	70.09	76.76	74.83	75.84
SiloBN [7]	50.83	63.81	53.98	61.90	57.63	83.17	81.41	77.90	80.56	80.76
IDA [21]	55.62	41.87	55.42	45.64	49.64	81.27	75.95	78.38	73.10	77.18
FedBN [14]	54.96	72.10	54.73	62.07	60.96	83.06	96.35	79.97	81.36	85.18
FedProto [16]	69.93	55.11	66.17	60.59	62.95	83.79	88.69	84.42	80.83	84.43
FML [5]	69.14	75.83	66.02	59.40	67.60	88.38	95.49	85.05	82.81	87.93
Ours	**71.92**	**85.33**	**68.39**	**65.63**	**72.82**	**89.08**	**98.16**	**85.47**	**84.11**	**89.20**

TABLE III
COMPARISON WITH STATE-OF-THE-ART FL WORKS ON REAL-WORLD PROSTATE MRI FL DATASET

Method	Dice (%)						IoU (%)					
	A	B	C	D	E	Avg	A	B	C	D	E	Avg
FedAvg [8]	71.21	77.58	68.77	52.76	75.25	69.11	67.12	73.18	64.97	50.42	70.67	65.27
FedProx [13]	70.38	78.22	68.04	54.89	79.87	70.28	67.25	74.14	64.33	51.03	**75.03**	66.36
SiloBN [7]	77.39	67.40	66.59	76.84	77.56	73.16	72.80	63.25	62.12	72.25	73.73	68.83
IDA [21]	74.72	80.09	66.80	57.38	74.21	70.64	70.79	75.41	63.41	53.23	71.83	66.93
FedBN [14]	74.57	75.21	65.69	76.10	79.72	74.26	69.16	70.11	60.59	71.99	74.35	69.24
FedProto [16]	74.91	78.91	68.87	78.93	**80.74**	76.47	69.52	73.12	64.11	72.21	74.14	70.62
FML [5]	**79.07**	80.84	70.17	79.65	74.85	76.92	**73.58**	74.33	65.11	72.76	69.54	71.06
Ours	77.66	**81.08**	**71.80**	**83.26**	77.61	**78.28**	71.29	**75.62**	**65.35**	**74.89**	72.46	**71.92**

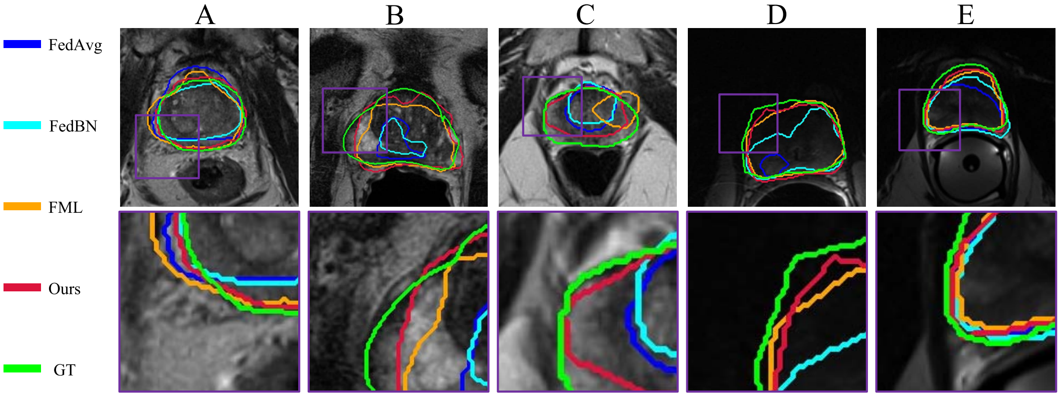

Fig. 4. Segmentation visualization of various FL approaches on five clients. The first row presents the segmentation results around prostate regions. The second row illustrates the details of segmentation contour within the purple box for better visualization.

2) Prostate MRI Segmentation: To comprehensively evaluate the performance of the proposed FL framework, we further perform the FL comparison on the prostate MRI segmentation dataset, as shown in Table III. Compared with the baseline FedAvg [8] with average Dice of 69.11% and average IoU of 65.27%, SiloBN [7] and FedBN [14] obtain a more than 4 percentage increase in average Dice, by alleviating the data heterogeneity among clients with specific normalization statistics. It is worth noting that FedProto [16] achieves better performance with average Dice of 76.47% by exploiting the abundant knowledge of segmentation prototypes. By utilizing PFA and DET to alleviate the retrogress problem and CPA loss to overcome the class imbalance in FL, our method achieves the outperforming segmentation results on most clients, with average Dice of 78.28% and average IoU of 71.92%. With five independent repeated experiments in Fig. 5 (b), our FL method significantly outperforms state-of-the-art FL works with the P-value <0.005 in both Dice and IoU.

Qualitative segmentation results of the prostate MRI FL dataset are elaborated in Fig. 4. These MRI slices reveal significant data heterogeneity among the samples of these five clients due to different imaging devices and protocols. To clearly demonstrate the segmentation results, we compare the proposed FL framework with the second-best FML [5], as well as the typical baselines FedAvg [8] and FedBN [14]. Compared with ground truth marked in green lines, our method achieves the most accurate prediction of prostate regions among state-of-the-art FL approaches. Especially on client B and client C, FedBN [14] and FML [5] are difficult to predict the presence of prostate at the marked regions, but our method better distinguishes the boundary between the prostate and backgrounds. These comparisons fully demonstrate that our method can achieve remarkable performance advantages in FL segmentation tasks from both quantitative and qualitative perspectives.

C. Ablation Study

1) Ablation of PFA, DET and CPA Loss: To investigate the effectiveness of the proposed PFA, DET and the CPA loss

TABLE IV
ABLATION STUDY ON REAL-WORLD DERMOSCOPIC FL DATASET

Row ID	PFA	DET	L_{CPA}		F1 (%)					AUC (%)				
			Global	Local	A	B	C	D	Avg	A	B	C	D	Avg
1					54.96	72.10	54.73	62.07	60.96	83.06	96.35	79.97	81.36	85.18
2	✓				65.30	69.64	65.51	61.97	65.60	83.77	95.41	83.20	82.41	86.20
3		✓			69.41	77.90	66.77	61.90	69.00	88.70	97.03	84.25	82.59	88.14
4			✓		68.16	67.00	67.10	55.12	64.35	85.13	87.16	83.90	80.07	84.07
5				✓	66.29	71.10	63.30	58.86	64.90	84.90	96.18	83.13	81.75	86.39
6			✓	✓	63.59	75.66	65.63	62.57	66.86	83.88	96.15	82.26	83.50	86.45
7	✓		✓	✓	67.42	74.24	66.30	61.57	67.39	85.47	97.70	83.47	81.21	86.96
8		✓	✓	✓	69.58	79.39	68.17	63.07	70.05	87.23	98.09	84.68	83.22	88.31
9	✓	✓			**72.06**	79.70	67.14	**65.82**	71.18	88.40	97.64	**86.04**	83.56	88.91
10	✓	✓	✓	✓	71.92	**85.33**	**68.39**	65.63	**72.82**	**89.08**	**98.16**	85.47	**84.11**	**89.20**

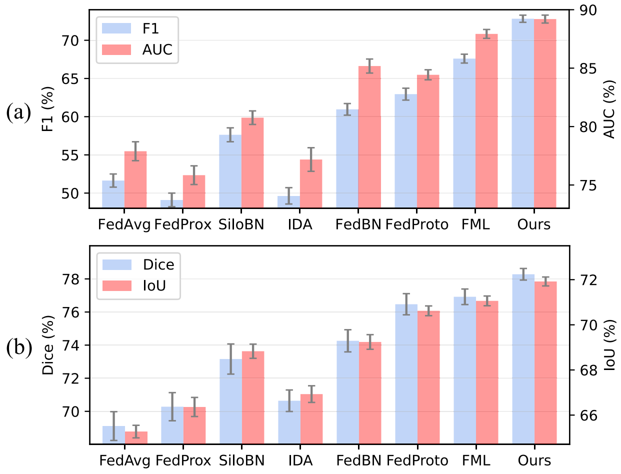

Fig. 5. Standard deviation of five independent repeated experiments for various FL works on (a) dermoscopic and (b) prostate MRI FL datasets.

TABLE V
ABLATION OF FREQUENCY COMPONENTS IN PFA

Amp.	Phase	Metrics	A	B	C	D	Avg
✓		F1	70.07	83.35	67.58	63.19	71.05
		AUC	87.54	96.98	85.52	84.12	88.54
	✓	F1	69.71	80.03	**69.68**	64.05	70.87
		AUC	87.57	96.28	85.12	**85.48**	88.61
✓	✓	F1	**71.92**	**85.33**	68.39	**65.63**	**72.82**
		AUC	**89.08**	**98.16**	**85.47**	84.11	**89.20**

TABLE VI
ABLATION STUDY OF THREE STEPS IN DET

E	R	S	Metrics	A	B	C	D	Avg
✓			F1	67.85	80.50	66.87	63.90	69.78
			AUC	87.52	97.82	84.68	82.72	88.18
✓	✓		F1	70.84	84.65	67.93	65.19	72.15
			AUC	88.71	97.93	84.95	84.07	88.92
✓	✓	✓	F1	**71.92**	**85.33**	**68.39**	**65.63**	**72.82**
			AUC	**89.08**	**98.16**	**85.47**	**84.11**	**89.20**

L_{CPA}, we further conduct the comprehensive ablation study on the dermoscopic FL dataset, as illustrated in Table IV. By removing the tailored modules from the proposed personalized retrogress-resilient FL framework, the 1^{st} row configuration (*i.e.*, FedBN [14]) serves as the baseline for the ablative comparison. By separately introducing the PFA (2^{nd} row) and DET (3^{rd} row), we observe that the baseline performance is improved with an individual F1 gain of 4.64% and 8.04%, respectively. These performance advantages can be attributed to the progressive aggregation of client knowledge in the frequency domain at the server side and the smoothed global knowledge transfer by introducing a deputy model at the client side. For the CPA loss, we investigate the effect of the global conjoint objective (4^{th} row) and the local prototype-aligned refinement (5^{th} row), which prove that these two improvements in CPA loss can alleviate the FL imbalance to a certain degree. By collaboratively adjusting the FL training from both global and local aspects, the proposed CPA loss (6^{th} row) can effectively promote the FL framework with a 5.90% F1 increase.

By comparing 7^{th} and 8^{th} rows with 2^{nd} and 3^{rd} rows, the CPA loss can further solve the FL imbalance on the basis of PFA and DET alleviating the retrogress problem, resulting in a F1 increase of 1.79% for the PFA case and 1.05% for the DET case. The PFA and DET employed at the server and client side (9^{th} row) can orthogonally overcome the bottleneck of existing FL frameworks, and result in superior performance among baselines, with the F1 of 71.18% and AUC of 88.91%. On this basis, our personalized retrogress-resilient FL framework (10^{th} row) further adopts the CPA loss to address the FL imbalance issue, which receives another F1 increase of 1.64%. In this way, these ablation experiments prove that the tailor-made PFA, DET and CPA loss play a significant role to address the real-world challenges in medical FL scenarios, leading to the performance improvement of the proposed FL framework.

2) Ablation of Frequency Components in PFA: In Table V, we investigate three possible combinations of frequency component aggregation in PFA, including amplitude-only components, phase-only components and both amplitude and phase components. These two baselines separately aggregate amplitude or phase component in PFA, together with the other original frequency component to construct the network parameters through IFFT. In our experiment, we observe that processing the amplitude-only components in PFA (average F1 of 71.05%) is slightly better than processing the phase-only components (average F1 of 70.87%) in terms of average F1. By comprehensively exploiting the useful knowledge in both amplitude and phase components of parameters, our framework obtains a 1.77% F1 increase over the amplitude-only baseline, which confirms the improvement in PFA compared to the conference version.

3) Ablation of Three Steps in DET: To analyze the advantage of our DET over the direct mutual learning [5], [15], we perform the ablation study of the three steps, including the Recover (R), Exchange (E) and Sublimate (S) steps. Compared

TABLE VII
ANALYSIS OF FREQUENCY THRESHOLDS r_0 AND r_1 IN PFA

r_0	0	0.10	0.20	0.30	**0.35**	0.40
F1 (%)	70.75	71.52	71.71	72.43	72.82	72.07
AUC (%)	88.30	88.99	88.73	89.10	89.20	88.38
r_1	0.40	0.42	0.44	0.46	**0.48**	0.50
F1 (%)	71.09	71.53	72.16	72.39	72.82	71.85
AUC (%)	88.51	88.52	88.82	89.08	89.20	88.90

TABLE VIII
ANALYSIS OF PERFORMANCE THRESHOLDS λ_1 AND λ_2 IN DET

λ_1	0.40	0.50	0.60	**0.70**	0.80
F1 (%)	70.88	71.44	72.06	72.82	72.04
AUC (%)	88.00	88.94	88.83	89.20	88.85
λ_2	0.75	0.80	0.85	**0.90**	0.95
F1 (%)	71.78	72.21	72.58	72.82	72.16
AUC (%)	88.62	88.40	88.34	89.20	88.61

TABLE IX
ANALYSIS OF HYPER-PARAMETERS β AND τ IN CPA LOSS

β	0.4	0.6	**0.8**	1.0	1.2
F1 (%)	71.95	72.26	72.82	72.05	71.69
AUC (%)	88.25	88.63	89.20	88.03	88.75
τ	1	2	**3**	4	5
F1 (%)	71.45	72.59	72.82	72.29	71.53
AUC (%)	87.87	89.18	89.20	88.51	88.72

with the direct Exchange step (average F1 of 69.78%) in Table VI, the preceding Recover step significantly improves the average F1 by 2.37%, by restoring the retrogress-affected deputy model to enable more efficient knowledge exchange for the personalized model. On this basis, the Sublimate step can further improve the average F1 by 0.67%, where the improved deputy model can focus on improving the personalized model while avoiding overfitting to the local data. As such, our DET significantly improves the direct mutual learning to focus more on improving the personalized model, by avoiding the impact of retrogress on the client side.

D. Analysis of Hyper-Parameters

As reported in Section IV-A, our FL framework involves six hyper-parameters, including frequency thresholds r_0 and r_1 in PFA, performance thresholds λ_1 and λ_2 in DET, and β and τ in CPA loss. Under the constraint as $0 \leq r_0 < r_1 \leq 0.5$ and $0 < \lambda_1 < \lambda_2 < 1$, we perform grid search to select these hyper-parameters in pairs, as presented in Table VII, VIII and IX. Overall, these comparisons confirm that the reported hyper-parameters are reasonable for our FL framework. In particular, the frequency thresholds r_0 and r_1, as well as performance threshold λ_1, are relatively more important for performance. Specifically, r_0 and r_1 determine the extent of parameter integration at the frequency space, and moderate values ensure that PFA can smoothly gather sufficient global knowledge at the server side. An appropriate λ_1 ensures the deputy model participating in DET with satisfactory performance, which avoids the damage of direct mutual learning to the personalized model. This confirms that the progressive strategy in PFA and the Recover step in DET are significant for our performance advantage. In contrast, our method is insensitive to changes of other hyper-parameters λ_2, β and τ, which proves that our FL framework is robust in practical applications.

TABLE X
GENERALIZATION COMPARISON ON OUT-OF-DISTRIBUTION DATA

Method	F1 (%)	AUC (%)
FedAvg [8]	37.76	66.49
FedProx [13]	38.14	67.36
SiloBN [7]	42.87	65.65
IDA [21]	40.11	62.75
FedBN [14]	43.74	70.85
FedProto [16]	40.68	65.97
FML [5]	37.01	66.81
Ours	**50.24**	**73.72**

E. Out-of-Distribution Generalization

To validate the generalization of the proposed FL framework on out-of-distribution cohort, we further evaluate the trained models on an unseen *ViDIR Legacy* cohort [12], with 350 *nv*, 10 *bkl* and 67 *mel* samples in Table I. For personalized FL methods, we select the local model of client A, because the class distribution of client A is closer to the whole FL framework. With more representative knowledge of the whole FL framework, the model of client A is relatively consistent with the global model in standard FL frameworks. As illustrated in Table X, our framework achieves the superior performance with 50.24% in F1 and 73.72% in AUC under such challenging scenario. Furthermore, our framework demonstrates remarkable performance advantages in comparison to existing FL approaches [5], [7], [8], [13], [14], [16], [21] with F1 increments of 12.48%, 12.10%, 7.37%, 10.13%, 6.50%, 9.56% and 13.23%, respectively. These comparisons confirm the superiority of our FL framework on model generalization.

V. DISCUSSION

In this section, we further investigate our FL framework from the perspective of class imbalance and personalized FL, and analyze the properties in terms of privacy, communication and computation.

A. Comparison With Imbalanced and Personalized FL

Beyond the FL works on data shift [5], [13], [14], [16], we perform a comparison with state-of-the-art FL imbalance studies [42], [43] (*i.e.*, targeted at label shift). Particularly, FedeAMC [42] utilized the number of local samples to re-balance the category-wise weight of cross entropy, and Fed-Focal [43] modified the loss based on the prediction accuracy to down-weight the well-classified samples of dominant categories. On the dermoscopic FL dataset, we adopt the same FL framework with PFA and DET (*i.e.*, 9^{th} row in Table IV), and individually apply the cross entropy loss L_{CE}, the CPA loss L_{CPA} and imbalance loss of FedeAMC [42] and Fed-Focal [43] as the L_{sup} in DET. As shown in Table XI, FedeAMC [42] and Fed-Focal [43] achieve average F1 of 71.23% and 71.97% respectively, which reveals a slight improvement over the baseline L_{CE}. Therefore, it is difficult to address the imbalance problem in medical FL by merely

TABLE XI
COMPARISON WITH FL IMBALANCE WORKS

Method	Metrics	A	B	C	D	Avg
L_{CE}	F1	**72.06**	79.70	67.14	**65.82**	71.18
	AUC	88.40	97.64	86.04	83.56	88.91
FedeAMC [42]	F1	71.06	80.31	68.28	65.28	71.23
	AUC	88.23	97.17	85.43	83.81	88.66
Fed-Focal [43]	F1	71.37	83.58	**70.43**	62.50	71.97
	AUC	88.17	97.63	**86.52**	83.71	89.01
L_{CPA} (Ours)	F1	71.92	**85.33**	68.39	65.63	**72.82**
	AUC	**89.08**	**98.16**	85.47	**84.11**	**89.20**

TABLE XII
COMPARISON WITH PERSONALIZED FL WORKS

Method	Metrics	A	B	C	D	Avg
Per-FedAvg [44]	F1	62.79	57.20	66.09	57.32	60.85
	AUC	81.50	89.92	81.18	82.67	83.82
pFedMe [45]	F1	63.84	60.97	63.09	59.79	61.92
	AUC	80.88	91.71	81.69	79.37	83.41
APFL [46]	F1	71.39	61.24	67.42	61.70	65.44
	AUC	86.60	91.35	84.40	80.54	85.72
Ours	F1	**71.92**	**85.33**	**68.39**	**65.63**	**72.82**
	AUC	**89.08**	**98.16**	**85.47**	**84.11**	**89.20**

adjusting the task loss of local training in terms of the number of local samples or the predicted probability. With the prototype-aligned refinement towards a global balanced goal, our L_{CPA} obtains a 1.64% increase in average F1 and achieves the superior F1 and AUC performance on most clients. This comparison proves the advantages of proposed CPA loss on the FL imbalance problem through the global and local collaborative refinement.

In addition to personalized SiloBN [7], FedBN [14] and FML [5], we further compare more personalized FL methods on dermoscopic FL dataset in Table XII, including Per-FedAvg [44], pFedMe [45] and APFL [46]. In general, compared with SiloBN [7] and FedBN [14] that benefit from personalized BN statistics, Per-FedAvg [44], pFedMe [45] and APFL [46] reveal impressive diagnostic performance with tailored personalization strategies for local optimization. In particular, APFL [46] achieves the average F1 of 65.44% by adaptively performing the local training of each client as the mixture of local model and global model. But for the most imbalanced client B, these FL methods obtain inferior performance to our method, which limits their average performance. In contrast, with the help of a deputy model, our FL framework ensures that personalized local model can focus on solving the local task, thereby leading to remarkable performance advantage. These comparisons confirm the advantage of our deputy strategy and tailored designs over advanced personalized FL works.

B. Privacy, Communication and Computational Analysis

Following standard FL frameworks [8], [14], our FL framework keeps the private data at the client side, and prevents the raw data from privacy leakage. Furthermore, with the progressive strategy in PFA, the server integrates the low-frequency components of the client parameters together, while excluding the high-frequency components that reflect the client's characteristics. When the aggregated models are delivered to corresponding clients, the risk of privacy leakage can be mitigated from the parameter attack at the client side.

We further take the dermoscopic FL diagnosis as an example to analyze the communication and computational cost. In each communication, each client in our FL framework would upload and receive one diagnostic model with the size of 9.10 MB, which is consistent with FedAvg [8], FedProx [13], IDA [21], SiloBN [7], FedBN [14] and FML [5]. In addition, our CPA loss demands to convey the prototypes between the server and clients, which is calculated by averaging the embeddings after the first FC layer with 64 output channels. Since each float scalar occupies 4 Bytes in PyTorch [36], the size of prototypes is 0.75 KB in the three-category diagnosis, as $3 \times 64 \times 4/1024$ KB. Compared with standard FL strategy, the communication overhead of prototypes is 0.008%, which is negligible in the FL framework.

For the computational cost, the prototypes merely demand 192 element-wise additions in each update. The computational overhead of prototype updating is negligible compared with the local model with computation of 1.96×10^9 FLOPs for each image. Except for the prototype updating, the computation of local training in our method is identical to FML [5], and 2 times of standard FL works [8], [14]. In medical FL, hospitals would serve as clients [4] and pursue superior diagnostic performance, where the cost of training the deputy model is completely affordable. When extended to the resource constrained FL (*e.g.*, edge computing), we can easily adopt the sparse training techniques [47] to significantly reduce the computation in the client-side local training, which overcomes the resource bottleneck for mobile devices.

VI. CONCLUSION

In this work, we systematically analyze retrogress and class imbalance issues in real-world medical FL, which severely degrades the performance of existing FL approaches. To tackle these two challenges, we propose a personalized retrogress-resilient FL framework from two perspectives of the server and clients. At the server side, we perform PFA to integrate the global knowledge from low-frequency to high-frequency gradually. At the client side, we utilize DET to transfer global knowledge by conducting three steps of Recover-Exchange-Sublimate, rather than replacing local models. Aiming at the class imbalance in FL, the CPA loss first estimates the global conjoint objective based on the sample counts, and then refines the local training of each client to eliminate the imbalance gap towards a balanced goal. Plenty of experiments are performed on real-world dermoscopic and prostate MRI FL datasets, which demonstrate the remarkable advantages of our FL framework over state-of-the-art approaches.

REFERENCES

[1] Z. Liu, R. Xiong, and T. Jiang, "Clinical-inspired network for skin lesion recognition," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2020, pp. 340–350.

[2] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[4] N. Rieke *et al.*, "The future of digital health with federated learning," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–7, Dec. 2020.

[5] T. Shen *et al.*, "Federated mutual learning," 2020, *arXiv:2006.16765*.

[6] H. R. Roth *et al.*, "Federated learning for breast density classification: A real-world implementation," in *Proc. MICCAI Workshop Distrib. Collaborative Learn.* Cham, Switzerland: Springer, 2020, pp. 181–191.

[7] M. Andreux, J. O. D. Terrail, C. Beguier, and E. W. Tramel, "Siloed federated learning for multi-centric histopathology datasets," in *Proc. MICCAI Workshop Distrib. Collaborative Learn.* Cham, Switzerland: Springer, 2020, pp. 129–139.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017, pp. 1273–1282.

[9] Z. Chen, M. Zhu, C. Yang, and Y. Yuan, "Personalized retrogress-resilient framework for real-world medical federated learning," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2021, pp. 347–356.

[10] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *Proc. ICLR*, 2020, pp. 1–16.

[11] Z. Liu, J. Xu, X. Peng, and R. Xiong, "Frequency-domain dynamic pruning for convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1051–1061.

[12] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, Dec. 2018.

[13] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.

[14] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in *Proc. ICLR*, 2021, pp. 1–27.

[15] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

[16] Y. Tan *et al.*, "FedProto: Federated prototype learning across heterogeneous clients," 2021, *arXiv:2105.00243*.

[17] M. J. Sheller *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Dec. 2020.

[18] W. Li *et al.*, "Privacy-preserving federated brain tumour segmentation," in *Proc. Int. Workshop Machine Learn. Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 133–141.

[19] Q. Dou *et al.*, "Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–11, Dec. 2021.

[20] D. Yang *et al.*, "Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101992.

[21] Y. Yeganeh, A. Farshad, N. Navab, and S. Albarqouni, "Inverse distance aggregation for federated learning with non-IID data," in *Proc. MICCAI Workshop Distrib. Collaborative Learn.* Cham, Switzerland: Springer, 2020, pp. 150–159.

[22] J. Tan *et al.*, "Equalization loss for long-tailed object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11662–11671.

[23] J. Wang *et al.*, "Seesaw loss for long-tailed instance segmentation," in *Proc. CVPR*, 2021, pp. 9695–9704.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[25] Y. Zhang, X.-S. Wei, B. Zhou, and J. Wu, "Bag of tricks for long-tailed visual recognition with deep convolutional neural networks," in *Proc. AAAI*, 2021, vol. 35, no. 4, pp. 3447–3455.

[26] B. Kang *et al.*, "Decoupling representation and classifier for long-tailed recognition," in *Proc. ICLR*, 2020, pp. 1–16.

[27] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9719–9728.

[28] H. Zheng *et al.*, "Refined local-imbalance-based weight for airway segmentation in CT," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2021, pp. 410–419.

[29] M. Frigo and S. G. Johnson, "FFTW: An adaptive software architecture for the FFT," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, May 1998, pp. 1381–1384.

[30] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[31] U. Michieli and M. Ozay, "Prototype guided federated learning of visual feature representations," 2021, *arXiv:2105.08982*.

[32] N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[33] N. Bloch *et al.*, "NCI-ISBI 2013 challenge: Automated segmentation of prostate structures," *Cancer Imag. Arch.*, 2015. [Online]. Available: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=21267207#21267207d170e52bc57d4c67b747b57bf88c460f

[34] G. Litjens *et al.*, "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Med. Image Anal.*, vol. 18, no. 2, pp. 359–373, 2014.

[35] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2713–2724, Sep. 2020.

[36] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[40] H. R. Roth *et al.*, "Federated whole prostate segmentation in MRI with personalized neural architectures," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2021, pp. 357–366.

[41] Z. Chen, J. Zhang, S. Che, J. Huang, X. Han, and Y. Yuan, "Diagnose like a pathologist: Weakly-supervised pathologist-tree network for slide-level immunohistochemical scoring," in *Proc. AAAI*, 2021, pp. 47–54.

[42] Y. Wang, G. Gui, H. Gacanin, B. Adebisi, H. Sari, and F. Adachi, "Federated learning for automatic modulation classification under class imbalance and varying noise condition," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 1, pp. 86–96, Mar. 2021.

[43] D. Sarkar, A. Narang, and S. Rai, "Fed-focal loss for imbalanced data classification in federated learning," in *Proc. IJCAI Workshop Federated Learn. Privacy Data Confidentiality*, 2020, pp. 1–7.

[44] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 3557–3568.

[45] C. T. Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with Moreau envelopes," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 21394–21405.

[46] Y. Deng, M. M. Kamani, and M. Mahdavi, "Adaptive personalized federated learning," 2020, *arXiv:2003.13461*.

[47] X. Qiu, J. Fernandez-Marques, P. P. Gusmao, Y. Gao, T. Parcollet, and N. D. Lane, "ZeroFL: Efficient on-device training for federated learning with local sparsity," in *Proc. ICLR*, 2021, pp. 1–16.