This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

pFLFE: Cross-silo Personalized Federated Learning via Feature Enhancement on Medical Image Segmentation

Luyuan Xie[1,2], Manqing Lin[1], Siyuan Liu[1,2], ChenMing Xu[1,2], Tianyu Luan[3], Cong Li[1,2], Yuejian Fang[1,2⋆], Qingni Shen[1,2⋆⋆], and Zhonghai Wu[1,2]

[1]School of Software and Microelectronics, Peking University, Beijing, China
[2]National Engineering Research Center for Software Engineering, Peking University, Beijing 100871, China
[3]State University of New York at Buffalo

Abstract. In medical image segmentation, personalized cross-silo federated learning (FL) is becoming popular for utilizing varied data across healthcare settings to overcome data scarcity and privacy concerns. However, existing methods often suffer from client drift, leading to inconsistent performance and delayed training. We propose a new framework, Personalized Federated Learning via Feature Enhancement (pFLFE), designed to mitigate these challenges. pFLFE consists of two main stages: feature enhancement and supervised learning. The first stage improves differentiation between foreground and background features, and the second uses these enhanced features for learning from segmentation masks. We also design an alternative training approach that requires fewer communication rounds without compromising segmentation quality, even with limited communication resources. Through experiments on three medical segmentation tasks, we demonstrate that pFLFE outperforms the state-of-the-art methods.

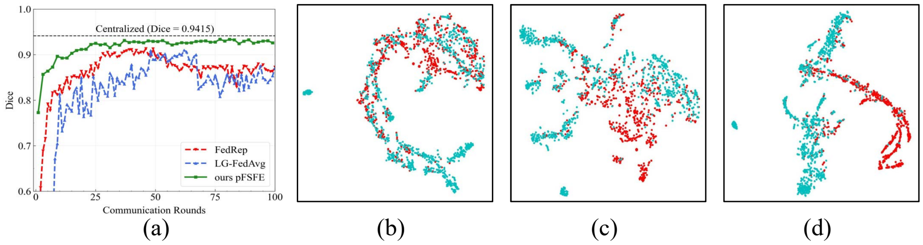
Keywords: Personalized cross-silo federated learning · Segmentation · Feature enhancement.

1 Introduction

Cross-silo Federated Learning (FL) [10,13,14,17] has recently achieved promising progress in medical image segmentation tasks such as [20,25,22]. It trains the network to gather data information from all clients without actually accessing data . Considering strict privacy regulation requirements of medical data, exploring cross-silo FL in medical image segmentation is important and useful to real-world applications [4,7,15,26]. To deal with data heterogeneous among clients caused by different medical protocols, personalized federated learning approaches such as [3,9,27,28] have made significant progress by using an aggregation of centralized and client-specific network design [18].

However, despite the recent progress made in personalized FL [5,8,16,12], previous works such as FedRep [1] and LG-FedAvg [2] still suffer from client drifting problems in medical image segmentation tasks [11,20,22]. The simple feature extraction

[⋆] Corresponding author: fangyj@ss.pku.edu.cn
[⋆⋆] Corresponding author: qingnishen@ss.pku.edu.cn



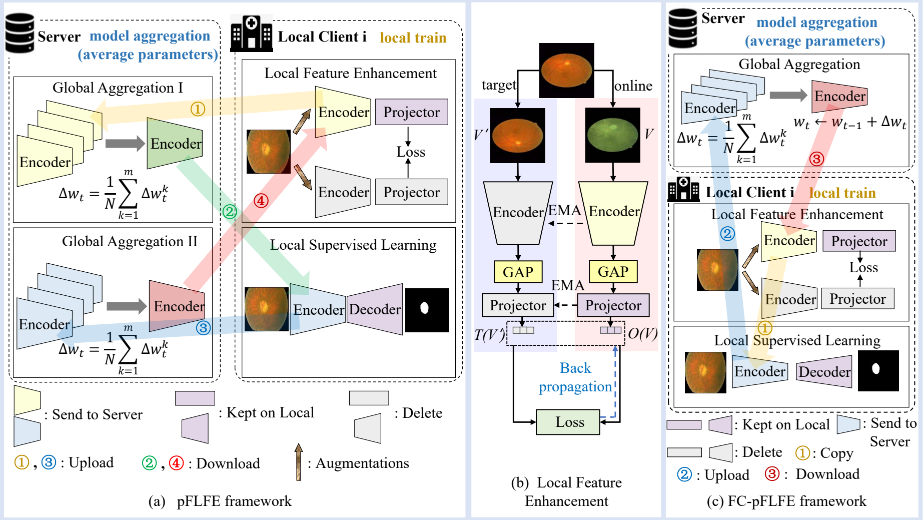
(a) (b) (c) (d)

Fig. 1: Previous personalized federated learning approaches suffer from client drifting problems. (a). Training progress of previous approaches (red and blue line) compared with our approach (green line). (b)(c) Foreground (cyan points) and background (red points) feature distribution t-SNE for FedRep and LG-FedAvg. (d) Feature distribution t-SNE for our pFLFE. The stability of previous works' training processes is not ideal and their foreground and background feature distribution are evidently overlapped. For comparison, pFLFE has a more stable training process, evidently split foreground and background feature distribution, and better segmentation accuracy.

and aggregation process would cause client drifting problems, which would degrade the performance and make the training process unstable. We show the training process of FedRep (red line) and LG-FedAvg (blue line) in Fig.1 (a). It is easy to observe that both FedRep and LG-FedAvg are not stable in training, and their performance drops as the training converges. To analyze the reasons that cause this issue, we visualize the t-distributed stochastic neighbor embedding (t-SNE) figure of foreground and background samples of FedRep and LG-FedAvg features in Fig.1 (b) and Fig.1 (c). The figures show clear distribution overlappings of foreground samples (cyan points) and background samples (red points) for both FedRep and LG-FedAvg. The indistinguishable feature distributions of foreground and background samples make the following network harder to classify foreground and background pixels.

To address these challenges, we introduce a new framework, namely personalized cross-silo Federated medical image Segmentation via Feature Enhancement (pFLFE), to enhance feature representation on each client without compromising data privacy. Our framework employs a self-supervised, contrastive approach for feature enhancement that relies solely on positive samples [30,19,31,32], reducing the need for large batch sizes and avoiding the sharing of features between clients [23,24,29]. This approach ensures data privacy and is particularly designed for medical applications where data is often limited.

Specifically, we design two federated learning frameworks including one with better performance and one that achieves comparable performance while using much fewer communication rounds. Our first framework consists of 4 main phases: local feature enhancement, local supervised training, and a global aggregation process after each local step, which together improve segmentation performance and training stability. The local supervised training will extract the information from each local client, and the local feature enhancement step will make personalized corrections to the shifted global parts after local training and aggregation. For our fast convergence framework, we use two similar local training steps but only one global aggregation process. The



(a) pFLFE framework

(b) Local Feature Enhancement

(c) FC-pFLFE framework

Fig. 2: The overview of (a) pFLFE framework, (b) Local Feature Enhancement and (c) FC-pFLFE framework.

simplified communication design is found to have similar segmentation performance while converging much faster.

In summary, our contributions include the following:

- We propose a novel personalized FL framework called pFLFE for medical image segmentation. pFLFE tackles client drift problems in medical image segmentation FL with a feature enhancement network using only positive samples, which eliminates the requirements of negative samples or features from other clients.
- We design an alternative fast-converging framework that can reach comparable performance in a few communication rounds, which is useful when communication resources are limited.
- Our experiments on 3 segmentation tasks involving in total of 17 datasets show that, pFLFE outperforms state-of-the-art results and achieves comparable performance with centralized learning, with high training stability and faster convergence.

2 Methods

2.1 pFLFE Framework

In Fig.2 (a), we present the pFLFE framework, consisting of 4 key stages: Local Feature Enhancement, Global Aggregation I, Local Supervised Learning, and Global Aggregation II, with stages 1 and 3 involving model uploads and stages 2 and 4 involving downloads. First, we enhance local feature extraction through a local feature enhancement module, which we will introduce details in Sec. 2.2. Then, during Global Aggregation I, encoders from all clients are aggregated. Subsequently, Local Supervised Learning refines segmentation networks on individual clients, extracting information from local data with annotation. Finally, Global Aggregation II shares refined encoders, further integrating supervised learning information across clients. In total, pFLFE uniquely improves local feature extraction and personalization, enabling superior performance and faster convergence without requiring data sharing, thus maintaining data privacy.

The components E_i, D_i, and P_i represent the encoder, decoder, and projector of the i-th client in Fig.2 (a), respectively. E_i is a shared parameter, and considering the issue of imbalanced client data and fairness in performance distribution [7,21], we adopt an average client weights approach rather than average data [21]. Due to the two-stage training of pFLFE, optimization losses for client i are as follows:

$$\begin{gathered}\mathcal{L}_{1,i} = \mathcal{L}_{MSE}\left(P_i(E_i(x_i^a)), P_i'(E_i'(x_i^{a'}))\right) + \mathcal{L}_{MSE}\left(P_i(E_i(x_i^{a'})), P_i'(E_i'(x_i^a))\right) \\ \mathcal{L}_{2,i} = \mathcal{L}_{Dice}(D_i(E_i(x_i)), y_i) + \mathcal{L}_{CE}(D_i(E_i(x_i)), y_i)\end{gathered} \tag{1}$$

$\mathcal{L}_{1,i}$ and $\mathcal{L}_{2,i}$ are the loss functions of the first and second stages, respectively. $\mathcal{L}_{MSE}$ means mean squared error loss, $\mathcal{L}_{CE}$ is cross-entropy loss and $\mathcal{L}_{Dice}$ is Dice loss [20]. $P_i'(\cdot)$ and $E_i'(\cdot)$ are another branch of client i in first stage. x_i and y_i are the data and labels on client i. a and a' are two different types of data augmentation.

2.2 Local Feature Enhancement

As shown in Fig.2 (b), the Local Feature Enhancement (LFE) consists of two branches: an online network O and a target network T. The target network is initialized with the same parameters as the online network. We train both branches using only positive samples, eliminating the need for negative samples or features from other clients. Given image I, we use two different data augmentation strategies for the same image. The resulting images are denoted as V and V', respectively. V and V' are passed through the encoder and projector in both the online and target branches, resulting in $O(V)$ and $T(V')$. In order for the encoder to learn the representation invariance of the same data, we calculate the mean squared error (MSE) between the normalized $O(V)$ and $T(V')$ like [**?**]:

$$\mathcal{L} = \left\| \overline{O(V)} - \overline{T(V')} \right\|_2^2 = 2 - 2 \cdot \frac{\langle O(V), T(V') \rangle}{\|O(V)\|_2 \cdot \|T(V')\|_2} \tag{2}$$

$\overline{O(V)}$ and $\overline{T(V')}$ are the ℓ_2-normalize of $O(V)$ and $T(V')$, respectively. $\langle \cdot \rangle$ means dot product, and $\|\cdot\|_2$ represents the ℓ_2-norm of the features. We separately feed V' to the online network and V to the target network to compute $\mathcal{L}' = \|\overline{O(V')} - \overline{T(V)}\|_2^2$, respectively. The total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L} + \mathcal{L}' \tag{3}$$

The target network updates through Exponential Moving Average (EMA) [23]. And we remove it after first training stage, which reduces the burden of local storage. In online branch, we upload encoder and retain the projector after the first training stage. Local Feature Enhancement can make personalized corrections to the previous round second stage of training and aggregated encoder that generates drift in each round.

Table 1: Test Dice coefficient comparison of optic disc/cup segmentation. "Client k Local" refers to the model training locally only using data on client k. We also report the $\text{Dice}_{\text{ACli}}$, $\text{Dice}_{\text{AImg}}$ and $\text{VDice}_{\text{ACli}}$. For $\text{Dice}_{\text{ACli}}$ and $\text{Dice}_{\text{AImg}}$, large is better. For $\text{VDice}_{\text{ACli}}$, smaller is better. **Bold** numbers indicate the best except for the centralized method. Underlines indicate the second bests. We can observe that our pFLFE and FC-pFLFE not only outperform SOTA in general, but also on most clients respectively. FT represents fine-tuning.

Model		Client1	Client2	Client3	Client4	Client5	Client6	Client7	$\text{Dice}_{\text{ACli}}\uparrow$	$\text{Dice}_{\text{AImg}}\uparrow$	$\text{VDice}_{\text{ACli}}\downarrow$
Centralized		0.9534	0.9425	0.9457	0.9532	0.9393	0.8987	0.9576	0.9415	0.9261	0.0185
Client1 Local		0.9172	0.4025	0.5314	0.0009	0.7076	0.3811	0.0446	0.4265	0.4698	0.3072
Client2 Local		0.7271	0.9001	0.8087	0.0002	0.7278	0.8218	0.5548	0.6486	0.7034	0.2830
Client3 Local		0.8483	0.8062	0.9054	0.0155	0.8524	0.8187	0.5174	0.6806	0.7428	0.2957
Client4 Local		0.0610	0.1299	0.0572	0.9322	0.0157	0.1299	0.4265	0.2503	0.1814	0.3055
Client5 Local		0.7904	0.6635	0.8329	0.0051	0.9153	0.7997	0.5529	0.6514	0.7358	0.2857
Client6 Local		0.8062	0.5368	0.3717	0.0037	0.5724	0.8815	0.6966	0.5527	0.6763	0.2746
Client7 Local		0.5465	0.8310	0.7339	0.0709	0.6579	0.6412	0.9242	0.6294	0.6113	0.2559
Single model	FedAvg	0.9226	0.8863	0.9274	0.2009	0.9121	0.8669	0.8404	0.7938	0.8247	0.2438
	FedAvg+FT	0.9379	0.9099	0.9246	0.5705	0.9037	0.8637	0.9171	0.8611	0.8617	0.1205
	SCAFFOLD	0.8970	0.8703	0.9157	0.4001	0.8993	0.8975	0.8372	0.8167	0.8487	0.1717
	SCAFFOLD+FT	0.9281	0.8833	0.9097	0.7855	0.8981	0.8796	0.8524	0.8767	0.8802	0.0433
	FedProx	0.9123	0.8824	0.8567	0.3301	0.8726	0.8821	0.8521	0.7983	0.8306	0.1920
	FedProx+FT	0.9301	0.8620	0.9273	0.7001	0.8801	0.8509	0.8743	0.8607	0.8567	0.0714
Presonalized FL	Ditto	0.6747	0.8830	0.9074	0.8335	0.8460	0.7350	0.9332	0.8304	0.7880	0.0869
	APFL	0.8762	0.8204	0.9338	0.8261	0.7253	0.7732	0.8341	0.8270	0.7873	0.0623
	LG-FedAvg	0.9530	0.8420	0.8917	0.9237	0.9234	0.6138	0.9486	0.8709	0.7909	0.1106
	FedRep	0.9342	0.9408	**0.9449**	0.7134	0.9161	0.8702	0.7873	0.8724	0.8751	0.0830
	FedSM	0.8995	0.8972	0.9216	0.8602	0.8251	0.8923	0.8659	0.8803	0.8733	0.0297
	LC-Fed	0.9173	**0.9422**	0.9149	0.8041	0.9022	0.8761	0.7992	0.8794	0.8805	0.0525
Ours	pFLFE	**0.9661**	0.9185	0.9317	**0.9381**	0.9300	**0.9002**	**0.9661**	**0.9358**	**0.9232**	**0.0222**
	FC-pFLFE	0.9633	0.8919	0.9072	0.9354	**0.9376**	0.8933	0.9630	0.9274	0.9194	0.0282

2.3 Fast Converging pFLFE

It is crucial to evaluate the communication overhead of federated learning. However, a complete training round of pFLFE includes two communication processes. This significantly increases the communication cost. To mitigate this issue, we provide a fast converging version of pFLFE named Fast Converging pFLFE (FC-pFLFE) as illustrated in Fig.2 (c), which can provide comparable results with very few communication rounds.

In the FC-pFLFE training phase, we begin with Local Feature Enhancement, which is the same feature enhancement procedure as in pFLFE. Then, the trained encoder is passed to the local segmentation model for Local Supervised Learning. After that, the encoder is then uploaded to the server for aggregation. FC-pFLFE only uses 1 aggregation step for 1 round of local feature enhancement and supervised learning, unlike pFLFE which uses 2 aggregation steps. This design allows us to reduce the number of communication rounds while largely preserving the feature enhancement and local knowledge acquisition of pFLFE.

3 Experiment Setup

Task and Dataset. We evaluate our proposed pFLFE on 3 tasks. **Optic Disc/Cup segmentation:** We use 7 datasets from [20,22]. Each dataset represents a client, so there are 7 clients. **Polyp segmentation:** The endoscopic images dataset is collected and annotated from four different centers [33,34,35,36], and each center's dataset is treated

Table 2: Test Dice coefficient comparison of polyp segmentation. FT represents fine-tuning.

Model		Client1	Client2	Client3	Client4	$\text{Dice}_{\text{ACli}}$↑	$\text{Dice}_{\text{AImg}}$↑	$\text{VDice}_{\text{ACli}}$↓
Centralized		0.8051	0.6910	0.8620	0.7765	0.7837	0.7977	0.0617
Client1 Local		0.7633	0.1775	0.3398	0.3119	0.3981	0.3861	0.2196
Client2 Local		0.2177	0.5332	0.2550	0.3744	0.3451	0.3280	0.1231
Client3 Local		0.2513	0.1212	0.8234	0.4208	0.4042	0.4771	0.2643
Client4 Local		0.3477	0.3012	0.5053	0.7357	0.4725	0.5649	0.1698
Single model	FedAvg	0.5249	0.4205	0.5676	0.5500	0.5158	0.5390	0.0570
	FedAvg+FT	0.6047	0.4762	0.7513	0.6681	0.6251	0.6632	0.1005
	SCAFFOLD	0.5244	0.3591	0.5935	0.5713	0.5121	0.5504	0.0918
	SCAFFOLD+FT	0.5937	0.4312	0.8231	0.7208	0.6422	0.7014	0.1464
	FedProx	0.5529	0.4674	0.5403	0.6301	0.5477	0.5770	0.0577
	FedProx+FT	0.7441	0.5701	0.7438	0.6402	0.6746	0.6809	0.0737
Presonalized FL	Ditto	0.5720	0.4644	0.6648	0.6416	0.5857	0.6201	0.0779
	APFL	0.6120	0.5095	0.6333	0.5892	0.5860	0.5984	**0.0468**
	LG-FedAvg	0.6053	0.5062	0.7371	0.5596	0.6021	0.6124	0.0855
	FedRep	0.5809	0.3106	0.7088	0.7023	0.5757	0.6479	0.1613
	FedSM	0.6894	0.6278	0.8021	0.7391	0.7146	0.7381	0.0641
	LC-Fed	0.6233	0.4982	0.8217	0.7654	0.6772	0.7325	0.1261
Ours	pFLFE	0.7895	**0.7447**	**0.8799**	**0.7705**	**0.7962**	**0.8021**	0.0509
	FC-pFLFE	**0.7962**	0.7182	0.8729	0.7653	0.7882	0.7965	0.0563

Table 3: Test Dice coefficient comparison of prostate segmentation. FT represents fine-tuning.

Model		Client1	Client2	Client3	Client4	Client5	Client6	$\text{Dice}_{\text{ACli}}$↑	$\text{Dice}_{\text{AImg}}$↑	$\text{VDice}_{\text{ACli}}$↓
Centralized Training		0.8157	0.8369	0.8678	0.8324	0.8469	0.7601	0.8266	0.8266	0.0336
Client1 Local		0.8062	0.0707	0.7680	0.0048	0.0520	0.0208	0.2871	0.2115	0.3544
Client2 Local		0.2585	0.7802	0.4863	0.0928	0.4498	0.4810	0.4248	0.4033	0.2128
Client3 Local		0.3342	0.0890	0.8045	0.0243	0.3769	0.1228	0.2920	0.2324	0.2624
Client4 Local		0.1760	0.1933	0.3115	0.6673	0.3448	0.2514	0.3241	0.3589	0.1647
Client5 Local		0.1062	0.5092	0.3118	0.4808	0.8488	0.4483	0.4509	0.4898	0.2239
Client6 Local		0.2853	0.3701	0.1764	0.0693	0.4684	0.5361	0.3176	0.2973	0.1612
Single model	FedAvg	0.5437	0.5881	0.7942	0.1412	0.7715	0.6364	0.5792	0.5244	0.2159
	FedAvg+FT	0.7373	0.7174	0.8274	0.5703	0.8569	0.6784	0.7313	0.7142	0.0948
	SCAFFOLD	0.5902	0.5312	0.7969	0.2888	0.7595	0.6349	0.6003	0.5548	0.1669
	SCAFFOLD+FT	0.6841	0.6607	0.8122	0.6423	0.8012	0.6533	0.7090	0.6980	0.0703
	FedProx	0.5261	0.4467	0.7521	0.1743	0.7640	0.5219	0.5309	0.4854	0.1990
	FedProx+FT	0.7745	0.7392	0.8321	0.5122	0.7823	0.7091	0.7249	0.6967	0.1024
Presonalized FL	Ditto	0.7677	0.7831	0.8385	0.6933	0.8395	0.3693	0.7152	0.7338	0.1624
	APFL	0.7233	0.8002	0.7143	0.5342	0.7976	0.5314	0.6835	0.6850	0.1115
	LG-FedAvg	0.7357	0.7252	0.7558	0.7130	0.8041	0.3431	0.6795	0.7053	0.1532
	Fedrep	0.7320	0.7517	0.7915	0.7050	0.7659	**0.7379**	0.7473	0.7384	**0.0272**
	FedSM	0.7643	0.8113	0.7503	0.7155	0.7859	0.6029	0.7384	0.7469	0.0674
	LC-Fed	0.7124	0.8259	0.8231	0.7752	0.8011	0.5124	0.7417	0.7592	0.1094
Ours	pFLFE	0.8300	**0.8604**	**0.8488**	0.7635	**0.8766**	0.6940	**0.8122**	**0.8145**	0.0639
	FC-pFLFE	**0.8449**	0.8279	0.8177	**0.8068**	0.8617	0.5848	0.7906	0.8055	0.0938

Table 4: Abation studies and the number of personalized layers. Only last layer: only use the last layer as the personalized part.

Method	Client1	Client2	Client3	Client4	$\text{Dice}_{\text{ACli}}$↑
pFLFE	**0.7895**	**0.7447**	**0.8799**	**0.7705**	**0.7962**
w/o LFE	0.7234	0.4373	0.8346	0.6954	0.6727
only last layer	0.7053	0.6154	0.7902	0.7043	0.7038

the number of personalized layers	Para(M)↓	$\text{Dice}_{\text{ACli}}$↑
only last layer	31.05	0.7038
last 3 layer	31.01	0.6744
last 9 layer	30.76	0.7893
last 14 layer	29.74	0.7341
last 19 layer	25.67	0.7654
All decoder	**18.85**	**0.7962**

Table 5: The KL divergence of personalized Federated learning framework with parameter decouple in three tasks. (LFE: Local Feature Enhancement)

Task	optic disc/cup	polyp	prostate	Average KL↓
FedRep	0.5012	0.7310	0.7710	0.6677
LG-FedAvg	0.4923	0.6512	0.8512	0.6649
LC-Fed	0.4631	0.6012	0.7412	0.6019
pFLFE (w/o LFE)	0.5542	0.7632	0.8932	0.7369
pFLFE	**0.4242**	**0.5227**	**0.5427**	**0.4965**

Table 6: The impact of different models on pFLFE in the polyp task

Model	Client1	Client2	Client3	Client4	$\text{Dice}_{\text{ACli}}$↑
FCN	0.7913	0.5056	0.6058	0.7717	0.7086
Unet	0.7895	0.7447	0.8799	0.7705	0.7962
Unet++	**0.8237**	**0.7563**	0.8854	**0.7926**	**0.8145**
Res-Unet	0.8103	0.7252	**0.8891**	0.7831	0.8019

as a separate client. This task has 4 clients. **Prostate segmentation:** We use MRI data collected and annotated by 6 institutions [11]. Each institution's dataset is treated as a client, resulting in 6 clients for this task. Considering that some clients have limited data, we employed a widely used 1:1 train-test split for optic disc/cup and polyp segmentation. The train-test split protocol follows the standard split used in the latest work for prostate segmentation.

Implementation Details. To ensure reliability of our experiments, we employ a five-fold cross-validation and report the mean Dice coefficient [20]. We calculate the average Dice coefficient for each client ($\text{Dice}_{\text{ACli}}$), the average Dice coefficient for all test images ($\text{Dice}_{\text{AImg}}$), and the variance of Dice across clients ($\text{VDice}_{\text{ACli}}$) to evaluate the model's performance and client discrepancy. (More details in supplementary materials.)

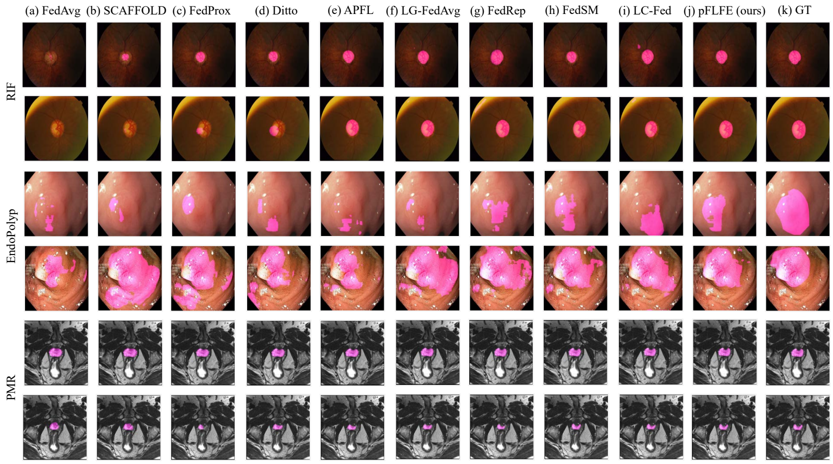


Fig. 3: Visualized comparison of personalized methods on three datasets. From each dataset, we randomly select two samples from different clients to form the visualization. (a-j) Segmentation results by models trained with FedAVG, SCAFFOLD, FedProx, Ditto, APFL, LG-FedAvg, FedRep, FedSM, LC-Fed and our method pFLFE; (k) Ground truths (denoted as 'GT');

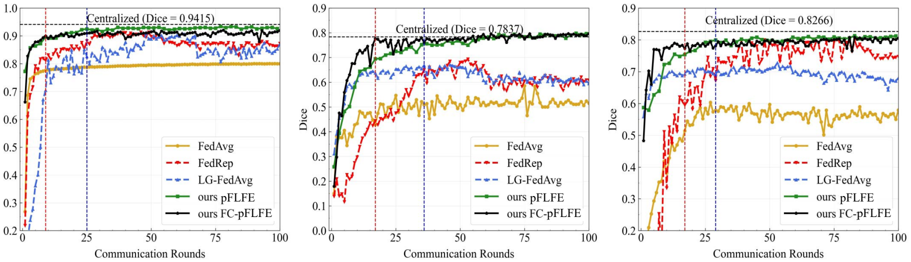
4 Results and Discussion

State-of-the-art comparison. Table 1, Table 2, and Table 3 present the comparisons between pFLFE and previous approaches on 3 tasks. We can easily observe that our pFLFE achieves optimal performance compared to other FL frameworks in all three tasks. And the performance of pFLFE is close to that of centralized learning, even surpassing centralized learning in polyp segmentation task. The FC-pFLFE performs slightly lower than pFLFE in all three tasks.

Ablation experiment. In Table 4, we conduct ablation experiments on pFLFE and adjusted the number of personalized layers of pFLFE. The results show that both LFE and the number of personalized layers have a significant impact on the performance of pFLFE. We use all decoders as personalized layers. This not only has performance advantages but also reduces the amount of parameters transmitted from the client to the server.

Feature distribution. In order to evaluate the quality of our enhanced feature, we compare the KL divergence of the feature distribution of the foreground and background samples. Smaller KL divergence indicates the global feature of foreground and background are better separated. As shown in Table.5, pFLFE feature has a smaller KL divergence than our baseline and previous works on all 3 tasks. These results verify the effectiveness of our feature enhancement design and the quality of our global feature.

The impact of different models. To demonstrate the impact of different models on pFLFE, we use different encoder-decoder segmentation networks on pFLFE. In Table 6, results show that using more complex models such as Unet++ and Res-Unet can help improve the performance of pFLFE.



(a) optic disc/cup segementation (b) polyp segementation (c) prostate segementation

Fig. 4: Training progress of our pFLFE compared with previous results on 3 tasks. The green and black line is our pFLFE and FC-pFLFE training progress, the red, blue, and yellow lines are training progress of previous approaches. The black dashed line is the result of the centralized method. It is easy to observe that our pFLFE has better performance and more stable, faster-converged training progress. The FC-pFLFE (red vertical line) reaches the near-optimal solution is significantly lower than that of pFLFE (blue vertical line).

Visualized comparison. Fig.3 visually compares the segmentation results on three tasks produced by our framework and baseline. Compared to other FL frameworks, pFLFE has better segmentation performance and is closer to ground truths.

The performance stability and convergence. Better performance stability and convergence are beneficial for subsequent model selection and reducing communication rounds. We compare FedAvg and other pFL frameworks (FedRep and LG-FedAvg) in Fig.4. The experimental results show that pFLFE has better convergence. During the communication process, pFLFE exhibits less performance fluctuations. This proves its good performance stability. Meanwhile, We compare FC-pFLFE with pFLFE. The results show that the number of communication rounds required for FC-pFLFE (red line) to reach the near-optimal solution is significantly lower than that of pFLFE (blue line).

Table 7: Comparison of federated domain generalization results on Optic Disc/Cup segmentation.

unseen	Client1	Client2	Client3	Client4	Client5	Client6	Client7	$\text{Dice}_{\text{ACli}}\uparrow$
FedAvg	0.9281	0.8999	0.9098	0.7888	0.8823	0.8257	0.8968	0.8759
SCAFFOLD	0.9127	0.8965	0.9122	**0.8855**	0.8394	0.8496	0.8923	0.8840
FedProx	0.9031	0.8445	0.9273	0.8701	0.8905	0.8042	0.8905	0.8757
Ditto	0.9099	**0.9173**	0.9221	0.8744	0.8523	0.8003	0.9056	0.8831
FedRep	0.9299	0.9013	0.9233	0.8723	0.8951	0.8423	0.8302	0.8849
ours	**0.9438**	0.9021	**0.9323**	0.8628	**0.9357**	**0.8806**	**0.9400**	**0.9139**

The domain adaptation capability of pFLFE. We verify the domain adaptation capability on Optic Disc/Cup segmentation tasks. In an unseen client, we extract and freeze their encoders trained on other clients, only fine-tuning the decoders. In Table 7, pFLFE obtains the optimal result. (More relevant experiments in supplementary materials.)

5 Conclusion

Our proposed pFLFE addresses features challenge faced in personalized cross-silo federated medical image segmentation. pFLFE incorporates Local Feature Enhancement and utilizes decoder as the personalized layer, ensuring both convergence and performance stability. We perform two-stage training and aggregation on global part (encoder), effectively improving its generalization and robustness. This enables it to adapt to new domains on unseen clients. Extensive experiments demonstrate superiority of our approach over existing personalized methods, approaching the performance of centralized learning.

Acknowledgements. This work was supported by the National Key R&D Program of China under Grant No.2022YFB2703301.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary. Federated learning with personalization layers. arXiv preprint arXiv:1912.00818, 2019.
2. Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. IEEE Intelligent Systems, 35(4):83– 93, 2020.
3. L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai. Exploiting shared representations for personalized federated learning. pages 2089–2099. PMLR, 2021.
4. F. Hanzely and P. Richtárik. Federated learning of a mixture of global and local models. arXiv preprint arXiv:2002.05516, 2020.
5. S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. pages 5132–5143. PMLR,2020.
6. Q. Li, B. He, and D. Song. Model-contrastive federated learning. In CVPR, pages 10713–10722, 2021.
7. T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. pages 6357–6368. PMLR, 2021.
8. T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems, 2:429–450, 2020.
9. P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdi- nov, and L.-P. Morency. Think locally, act globally: Federated learning with local and global representations. arXiv preprint arXiv:2001.01523, 2020.
10. Xie, L., Lin, M., Luan, T., Li, C., Fang, Y., Shen, Q., and Wu, Z. Mh-pflid: Model heterogeneous personalized federated learning via injection and distillation for medical data analysis. *arXiv preprint arXiv:2405.06822*.
11. Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In CVPR, pages 1013–1023, 2021.
12. Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh. Three approaches for personalization with applications to federated learning. arXiv preprint arXiv:2002.10619,2020.

13. O. Marfoq, G. Neglia, R. Vidal, and L. Kameni. Personalized federated learning through local memorization. pages 15070–15092. PMLR, 2022.
14. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. pages 1273–1282. PMLR, 2017.
15. J. Mills, J. Hu, and G. Min. Multi-task federated learning for personalised deep neural networks in edge computing. IEEE Transactions on Parallel and Distributed Systems, 33(3):630–641, 2021.
16. F. Sattler, K.-R. Müller, and W. Samek. Clustered federated learning: Modelagnostic distributed multitask optimization under privacy constraints. IEEE transactions on neural networks and learning systems, 32(8):3710–3722, 2020.
17. Y. Tan, G. Long, J. Ma, L. Liu, T. Zhou, and J. Jiang. Federated learning from pretrained models: A contrastive learning approach. Advances in Neural Information Processing Systems, 35:19332–19344, 2022.
18. J. Wang, Y. Jin, and L. Wang. Personalizing federated medical image segmentation via local calibration. In ECCV, pages 456–472. Springer, 2022.
19. Y. Wu, D. Zeng, Z. Wang, Y. Sheng, L. Yang, A. J. James, Y. Shi, and J. Hu. Federated self-supervised contrastive learning and masked autoencoder for dermatological disease diagnosis. arXiv preprint arXiv:2208.11278, 2022.
20. A. Xu, W. Li, P. Guo, D. Yang, H. R. Roth, A. Hatamizadeh, C. Zhao, D. Xu, H. Huang, and Z. Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In CVPR, pages 20866–20875, 2022.
21. J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research, 5:1–19, 2021.
22. L. Yi, J. Zhang, R. Zhang, J. Shi, G. Wang, and X. Liu. Su-net: an efficient encoder-decoder model of federated learning for brain tumor segmentation. pages 761–773. Springer, 2020.
23. Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33:21271–21284, 2020.
24. Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. pages 2–25. PMLR, 2022.
25. Li, W.; Milletar‘ı, F.; Xu, D.; Rieke, N.; Hancox, J.;Zhu, W.; Baust, M.; Cheng, Y.; Ourselin, S.; Cardoso, M. J.; et al. 2019. Privacy-preserving federated brain tumour segmentation. In MLMI, 133–141. Springer.
26. Huang, Y.; Chu, L.; Zhou, Z.; Wang, L.; Liu, J.; Pei, J.; and Zhang, Y. 2021. Personalized cross-silo federated learning on non-iid data. In AAAI, volume 35, 7865–7873.
27. Yang, D.; Xu, Z.; Li, W.; Myronenko, A.; Roth, H. R.; Harmon, S.; Xu, S.; Turkbey, B.; Turkbey, E.; Wang, X.; et al. 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multinational data from China, Italy, Japan. Medical image analysis, 70: 101992.
28. Roth, H. R.; Yang, D.; Li, W.; Myronenko, A.; Zhu, W.; Xu, Z.; Wang, X.; and Xu, D. 2021. Federated whole prostate segmentation in MRI with personalized neural architectures. In MICCAI, 357–366. Springer.
29. Qi, X.; Yang, G.; He, Y.; Liu, W.; Islam, A.; and Li, S. 2022. Contrastive re-localization and history distillation in federated CMR segmentation. In MICCAI, 256–265. Springer.
30. Xie, L., Li, C., Wang, Z., Zhang, X., Chen, B., Shen, Q., and Wu, Z. Shisrcnet: Super-resolution and classification network for low-resolution breast cancer histopathology image, 2023.
31. Dong, N.; and Voiculescu, I. 2021. Federated contrastive learning for decentralized unlabeled medical images. In MICCAI, 378–387. Springer.

32. Wu, Y.; Zeng, D.; Wang, Z.; Shi, Y.; and Hu, J. 2022. Distributed contrastive learning for medical image segmentation. Medical Image Analysis, 81: 102564.
33. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *IJCARS*, volume 9, pp. 283–293. 2014.
34. J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE TMI*, volume 35, pp. 630–644. 2015.
35. J. Bernal, F. J. S´ anchez, G. Fern´andez-Esparrach, D. Gil, C. Rodrıguez, and F. Vilarino Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG*, volume 43, pp. 99–111, 2015.
36. D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen Kvasir-seg: A segmented polyp dataset. *MMM*, 2020.
37. Xie, L., Li, C., Zhang, X., Zhai, S., Fang, Y., Shen, Q., and Wu, Z. Trls: A time series representation learning framework via spectrogram for medical signal processing, 2024.