IEEE Transactions on Medical Imaging

Personalized Federated Medical Image Segmentation via Virtual Classes

Journal:	*IEEE Transactions on Medical Imaging*
Manuscript ID	TMI-2024-2053
Manuscript Type:	Regular Paper
Date Submitted by the Author:	12-Aug-2024
Complete List of Authors:	Liu, Yuzhi; Shenzhen University, College of Computer Science and Software Engineering Fang, Zhixue; Shenzhen University, College of Computer Science and Software Engineering Wu, Huisi; Shenzhen University, College of Computer Science and Software Engineering; Qin, Jing; The Hong Kong Polytechnic University, Centre for Smart Health, School of Nursing
Keywords:	Federated Learning < General methodology, Segmentation < General methodology, Eye < Object of interest, Decentralized Learning < General methodology

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. XX, XXXX 2024

Personalized Federated Medical Image Segmentation via Virtual Classes

Yuzhi Liu, Zhixue Fang, Huisi Wu, *Senior Member, IEEE*, and Jing Qin, *Senior Member, IEEE*

***Abstract*— Federated learning shows promise for collaborative privacy-preserving medical image segmentation by aggregating parameters from diverse sites. Nevertheless, a single global model struggles with statistical heterogeneity across healthcare institutions. Personalized federated learning offers an effective solution by customizing models locally. However, data heterogeneity issues become exacerbated as client models diverge. Existing methods to mitigate these challenges have limitations in balancing global semantic consistency and local personalization, privacy protection, and segmentation effectiveness. To address this, we propose a novel virtual class strategy, which allows pixel features to be personalized based on global standards. Virtual classes are additional classes added to the classification head and frozen during training as consistent standards. They explicitly refine the same class across heterogeneous clients, alleviating the limitations of representing diverse clients with a single class. Additionally, we have developed a personalized affinity aggregation scheme to prevent damage to client-specific recognition of virtual classes, which is caused by high-resource clients during parameter aggregation. Experiments on two medical image datasets validate substantial performance gains, demonstrating the efficacy of our proposed techniques in advancing robust personalized federated segmentation. Codes will be released upon publication.**

***Index Terms*— Personalized Federated Learning, Medical Image Segmentation, Virtual Class**

I. Introduction

FEDERATED learning (FL) has emerged as a privacy-preserving approach for training machine learning models using diverse datasets from various sources [1], [2]. It shows significant promise for sensitive medical applications that demand strict confidentiality of healthcare data [3], [4]. However, a single global model is difficult to perform well on all healthcare institutions due to data heterogeneity, which is caused by variations in imaging devices, acquisition protocols, and patient populations [5], [6]. To address this challenge, personalized federated learning (PFL) has emerged as a promising solution [7]. PFL aims to tailor an optimal model for each client based on their local imaging data, thus allowing for model adaptation to institution-specific demographics and use cases while preserving data privacy [8]. However, the challenges caused by heterogeneity become more severe in PFL as client models diverge across sites. As illustrated in Fig. 1, an ophthalmology clinic can locally distinguish the optic disc from the cup. However, some local features of the optic disc may be more similar to the optic cup features of other sites due to statistical heterogeneity. Consequently, the disc features may drift towards the cup features from other sites after aggregation. This semantic ambiguity arises because a single global classifier struggles to make optimal decisions for heterogeneous site data.

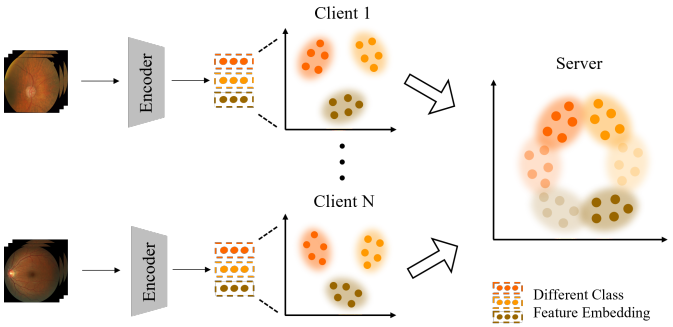

Fig. 1: In personalized federated learning, disparate data distributions and limited data accessibility across sites lead to significant feature overlap during global aggregation. This overlap introduces ambiguity in the global semantic representation, hindering the learning process at each client site.

Although some personalization methods like parameter decoupling [9], model interpolation [10], multi-task [11], meta-learning [12] and personalized network layers [13] help mitigate the challenges of data heterogeneity, they focus more on personalized feature extraction capabilities for image classification tasks. There are also some methods that seek to achieve consistent feature learning by globally sharing categorical feature centers [9] or fixed global classifier [14]. However, relying solely on a single shared feature centre per category cannot satisfy all clients' needs. Additionally, the effectiveness of these works for fine-grained pixel segmentation tasks remains to be validated, as they lack personalized components tailored for fine-grained segmentation. Recent personalized segmentation works [15], [16] have proposed learning from inter-site inconsistencies to improve local personalization capabilities. Nevertheless, over-specializing models to local distributions can compromise cross-site generalization. Moreover, directly exchanging information across sites to learn inconsistencies

Y. Liu, Z. Fang, and H. Wu are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (Email: hswu@szu.edu.cn).

J. Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Kowloon, Hong Kong (Email: harry.qin@polyu.edu.hk).

can infringe on privacy and increase computational costs. Consequently, existing methods struggle to strike a balanced trade-off between local personalization and global generalization.

To address these issues, we propose a novel strategy that enables clients to achieve personalization based on consistent global patterns. The core concept of this strategy is the *virtual classes*, which are additional classes added to the original classification head and frozen during the training as consistent learning standards for the client. We orthogonally initialize the classification head to broadly cover the classification space, facilitating personalization capture across heterogeneous site data. Specifically, in addition to the real anatomical class assignments, we adaptively allocate each pixel to one of the virtual classes based on maximum feature similarity. This adaptive allocation enables the personalized of pixels while aligning them with the global standard represented by the frozen virtual class. These virtual classes can be considered as supplementary discriminants for the real classes, mitigating the limitations of using a single category description across heterogeneous sites.

However, the personalized capabilities of clients tend to favor those with more data during parameter aggregation. This bias leads to the improper assignment of pixels to virtual classes in limited data clients. Hence, we propose a personalized affinity aggregation module to selectively filter knowledge from the global context that aligns with local adaptations, enhancing the ability of local virtual assignments. This module comprises decoder affinity queries, encoder affinity keys, and aggregation matrices to integrate global feature extraction and localized semantics decoding capabilities. During client parameter updates, we locally update decoder query and matrix parameters to enable client-specific learning. Simultaneously, clients aggregate encoder key parameters to retain universal feature extraction capabilities. By modeling interactions between global and local contexts, clients can better preserve personalized virtual allocation capabilities while leveraging shared knowledge. Overall, our main contributions can be summarized as follows:

- We propose a virtual class strategy aimed at enriching class representations to meet the needs of heterogeneous sites. Virtual classes can be seen as diverse global learning standards, allowing sites to fine-tune personalized adjustments based on this standard.
- We develop a personalized affinity aggregation module to prevent the domination of virtual class personalization allocation by multi-resource clients. It allows the client to maintain its personalized virtual allocation while benefiting from global knowledge.
- Our proposed approach demonstrated substantial performance gains over state-of-the-art PFL techniques across various experimental configurations on two real-world medical datasets.

II. Related Work

A. Federated Learning

Federated learning has gained increasing research attention as a distributed learning framework that preserves data privacy [1], [2]. It learns a shared global model by aggregating model parameters from clients while keeping each client's data localized and not directly shared [17]. However, statistical heterogeneity across diverse client distributions remains a crucial challenge. A seminal work is FedAvg [18], which proposes aggregating site parameters weighted by local data size to mitigate this challenge. Additionally, many solutions tackle heterogeneity in federated learning, mainly following two paradigms: based on improving localized training [19], [20] or optimizing centralized aggregation [21], [22].

B. Personalized Federated Learning

Traditional federated learning aggregates model updates from heterogeneous clients to train a globally applicable model. However, statistical differences across sites mean this single global model often cannot meet the needs of diverse clients [23]. Personalized federated learning overcomes this limitation by learning customized local models tailored to each institution's data characteristics [24]. Paradigms like meta-learning [12], transfer learning [11], local fine-tuning [25], [26], and model interpolation [7], [10] have shown promise for classification. Besides, some methods [9] propose the global sharing of categorical prototypes for consistent features. However, a single prototype per class struggles to satisfy diverse clients and risks privacy. Furthermore, their efficacy for fine-grained medical image segmentation remains less explored. Recent methods [15], [16] aim to utilize cross-client inconsistencies for learning. However, excessively focusing on local information for personalization would compromise global generalization. Moreover, these methods encounter challenges in terms of computational efficiency and privacy protection. To alleviate this issue, we propose the use of virtual classes to complement the challenges of describing heterogeneous data using a single category. They serve as fixed standards in the global context, mitigating the impact of heterogeneous data while capturing site-specific personalization.

C. Virtual Classes

Virtual classes represent a set of labels unrelated to real classes. It has been empirically demonstrated that this approach is effective in reinforcing decision boundaries and compressing intra-class distributions [27]. Recent works in incremental learning used virtual classes to reserve feature space for future classes [28], [29]. Our crucial novelty is introducing virtual class into federated learning to mitigate semantic drift across heterogeneous data while achieving personalization.

III. Method

Our proposed personalization federated learning framework is illustrated in Fig. 2. To address the limitations of single-class representations for heterogeneous data, we introduce virtual classes to enrich the representation of heterogeneous data within the same class. We also design a personalized affinity aggregation to maintain site-specific virtual classes relevant to their own data while incorporating global knowledge. More details will be discussed in the following sections.

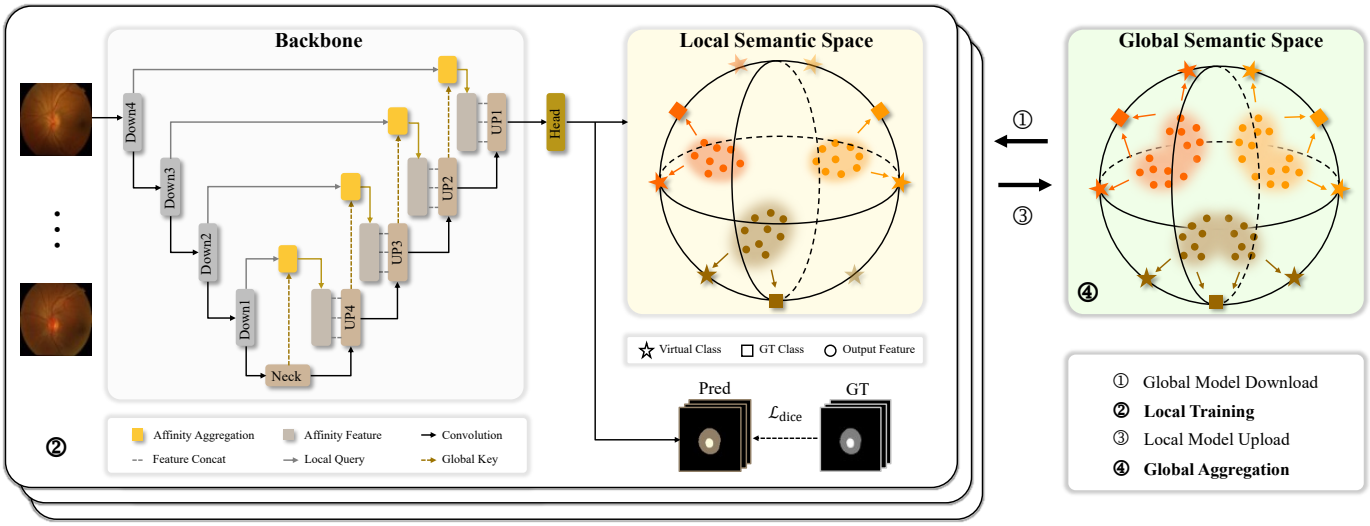

Fig. 2: An overview of our proposed framework. Our main contributions involve the local training and global aggregation components. Specifically, we introduce virtual classes and personalized affinity aggregation to enhance personalization while maintaining inter-client consistency.

A. Personalized Federated Learning Paradigm

Suppose there are N sites with non-independent and identically distributed data denoted $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^{N}$. Each site can access their S_i samples in the form of $\left(x_j^i, y_j^i\right)_{j=1}^{S_i} \sim \mathcal{D}_i$. In personalized federated segmentation, our goal is to learn N segmentation models denoted as $\{\theta_1, \theta_2, \cdots, \theta_N\}$, which are customized for local data distribution of each site:

$$\min_{\theta_1,\cdots,\theta_N} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i(\theta_i; \mathcal{D}_i) \tag{1}$$

$$\mathcal{L}_i(\theta_i; \mathcal{D}_i) = \frac{1}{S_i} \sum_{j=1}^{S_i} \mathcal{L}_i\left(f_i\left(x_j^i\right), y_j^i\right) \tag{2}$$

where f_i is the i^{th} local model and $\mathcal{L}_i$ is the segmentation loss function.

B. Virtual Class in Personalized Federated Learning

Previous work [14] has shown that fixing global classification heads provides consistent learning standards and guidance for each category. However, relying on a single representation struggles to accommodate heterogeneous data of the same class across sites. It also limits the fine-grained personalization needed for segmentation tasks.

Therefore, we propose utilizing additional virtual classes to provide multiple fixed yet flexible learning standards for each category. Fixed involves inserting a specific number of frozen virtual classes globally, while flexible allows clients to adaptively choose virtual classes based on the local data. Specifically, we introduce extra virtual classes in the classification head as supplementary fixed classifiers to enable consistent intra-class learning guidance, as illustrated in Fig. 3 (a) and (b). In other words, we adaptively assign each pixel a globally fixed virtual class while learning the true labels. This allocation can be viewed as a personalized process, enabling sites to fine-tune within classes based on their data features. As shown in Fig. 3 (c), heterogeneous data features from the same real class are drawn closer to the same virtual class while maintaining their real label, whereas dissimilar same-class features are assigned to distinct virtual classes. In this way, the multiple virtual classes empower more fine-grained and tailored learning compared to relying on a single representation.

A key question is how to design virtual classes and enable adaptive pixel-to-virtual class associations. We observe that in prediction heads of the network, the convolutional computation between feature maps and kernels can be approximately viewed as a similarity calculation between feature maps and kernel weights. This perspective of similarity-based parsing provides a solution: allocate each pixel to the most compatible virtual class based on maximal feature-class affinity.

Specifically, assuming we decode a feature map with dimensions $H \times W \times C_1$, representing the image dimensions and feature channels. This feature map is processed through a convolutional head to produce an output probability map of $H \times W \times C_2$, where C_2 represents the number of classes. The convolutional head uses weights of $(C_2, C_1, K \times K)$, where K is the kernel size. When $K = 1$ and stride $= 1$, the kernel slides over each pixel of the feature map. It performs a dot product operation between the corresponding feature map C_1 channels and its C_1 weights, generating a new similarity value. With C_2 such kernels, each pixel position ultimately yields C_2 similarity values.

Hence, we define the virtual classes as the parameters of the prediction head convolution kernels. We orthogonally initialize the parameters of kernels to ensure uncorrelated virtual concepts, which can capture the heterogeneous data from various sites with diverse distributions. Critically, the virtual classes remain frozen, preventing them from being distorted by personalized adaptations for each client.

Technically, we decouple the model f into a base network g and a segmentation head h. For an input sample x, the model can be expressed as $f(x) = h(g(x))$. We parameterize the last segmentation head layer h into real class and virtual

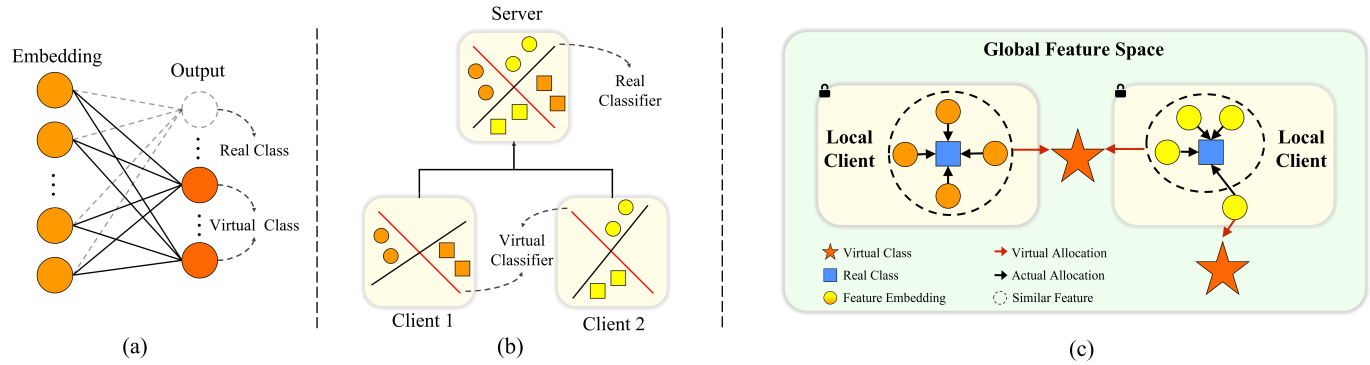

Fig. 3: Illustration of virtual classes. (a) Virtual classes are performed by adding additional frozen classes to the classifier header. (b) Virtual classes can are supplement to real classes, capturing the personalized needs of heterogeneous sites. (c) We perform bimodal allocation for pixels: the real class and the globally fixed virtual class.

class and classify pixels based on the similarity between pixel embeddings and the classes:

$$f(x) = \text{Conv}\left(\frac{Z}{\|Z\|_2}; \frac{\theta_h}{\|\theta_h\|_2}\right) \tag{3}$$

where θ_h denotes the parameters of the last convolutional layer of the segmentation head, and $Z = g(x)$ is the pixel embeddings from the base network for input x. Here, we first apply regularization on θ_h and Z to mitigate scale discrepancies, and then use 1×1 convolutions to compute semantic similarity scores between the regularized pixel embeddings Z and regularized class weights θ_h.

Thus, we obtain $C + V$ segmentation maps instead of only traditional C maps, where C is the number of real classes and V is the number of virtual classes. As a result, the pixel real class cross-entropy loss function $\mathcal{L}_r$, and pixel virtual class cross-entropy loss function $\mathcal{L}_v$ could be represented as follows:

$$\mathcal{L}_r = -\frac{1}{P}\sum_{i=1}^{P}\sum_{c=0}^{C-1} y_{i,c}\log\left(s_{i,c}\right) \tag{4}$$

$$\mathcal{L}_v = -\frac{1}{P}\sum_{i=1}^{P}\sum_{c=C}^{C+V-1} \hat{y}_{i,c}\log\left(\hat{s}_{i,c}\right) \tag{5}$$

where P represents the total number of sample pixels. $s_{i,c}$ denotes the similarity between the pixel and the real classes, and $\hat{s}_{i,c}$ denotes the similarity between the pixel and the virtual class. $\hat{y} = \arg\max\left(\hat{s}_{i,c}\right)$ is the pseudo label of the virtual class obtained from the maximum similarity. Note that we explicitly mask the output for the real label to avoid damaging the real prediction when computing the virtual class loss. Formally, it can be expressed as follows:

$$\hat{s}_{i,c} = \mathcal{M}_{i,c}\left(s_{i,c}\right) \tag{6}$$

$$\mathcal{M}_{i,c}\left(s_{i,c}\right) = \begin{cases} s_{i,c} & C \le c < V \\ 0 & 0 \le c < C. \end{cases} \tag{7}$$

We enforce a dual-peaked output for each pixel, assigning it to both the ground truth class and the nearest virtual class fixed in the global semantic space. By learning its ground truth class, the model preserves the ability to capture actual granular semantics for segmentation based on each client's unique data characteristics. Simultaneously, mapping pixels to the closest virtual class in a shared global semantic space mitigates harmful divergence across distributed client models. In summary, our loss function is designed as follows:

$$\mathcal{L} = \mathcal{L}_{dice} + \lambda_r\mathcal{L}_r + \lambda_v\mathcal{L}_v \tag{8}$$

where $\mathcal{L}_{dice}$ is dice loss function, λ_r and λ_v is the balance factor between real and virtual classes.

Why virtual classes work: In Eq.4, we only optimize gradients from the original C classes, aligning with traditional segmentation. This ensures that pixels remain real to their class while avoiding non-target classes. Eq.5 only focuses on gradients from virtual classes, aiming for pixels to choose the closest virtual class based on output features. We mask the logits of real classes (in Eq.5) to prevent them from being optimized by the learning of virtual classes. Besides, virtual class (added convolution kernel weight) should be frozen to maintain a consistent and unbiased standard across all iterations and clients. Other parameters are optimized normally. Our method of separately calculating real and virtual class losses allows each pixel to be classified by both, helping the model handle complex classes across clients.

C. Personalized Affinity Aggregation

While the proposed virtual class module can substantially reduce semantic divergences induced by federated aggregation, inherent limitations remain under FedAvg aggregation [18]. By simply averaging client model updates, FedAvg [18] allows high-resource contributors to dominate the global model direction. Such homogenizing effects hamper the personalization capability, as pixels from smaller, unique local datasets may be improperly mapped to these unrepresentative virtual classes during training.

To address this, we aim to enable clients to retain personalized capabilities for their virtual allocations based on their own data while integrating global knowledge. Prior works [30], [31] have shown that fusing encoder information during decoding benefits fine-grained discrimination. Inspired by this, we inject hierarchical affinity aggregation throughout the network encoder and decoder to selectively fuse encoder and decoder features suitable for local data, which enhances the local site's capability to identify virtual classes individually.

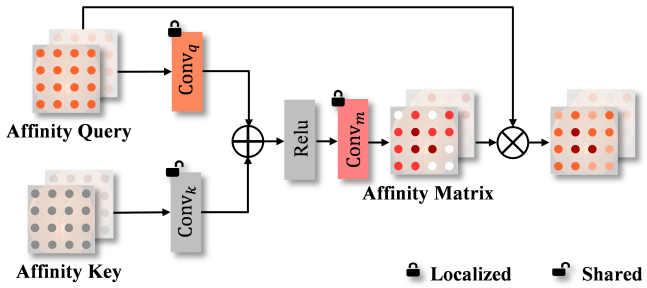

Fig. 4: Workflow of the personalized affinity aggregation. The affinity query from the decoder is localized along with the affinity aggregation matrix, while the affinity key from the encoder is globally shared.

Specifically, we consider a U-Net with L layers. Let E_{l-1} and E_l denote the encoder feature of the $(l-1)^{th}$ and l^{th} layer, respectively. Similarly, D_{l-1} and D_l are the decoder features for the $(l-1)^{th}$ and l^{th} layer, respectively. To compute the l^{th} affinity matrix, we first obtain the affinity query Q from D_l and the affinity key K from E_{l-1}. For ease of presentation, we ignore the superscript l of the Q and K at layer l. Formally, the affinity query and affinity key can be expressed as:

$$Q = \text{Conv}_q\left(D_l\right) \tag{9}$$

$$K = \text{Conv}_k\left(E_{l-1}\right) \tag{10}$$

where Conv_q and Conv_k are the feature projection operation. Hence, the affinity matrix M is computed:

$$M = \text{Conv}_m\left(\sigma\left(Q \oplus K\right)\right) \tag{11}$$

where Conv_m is the affinity matrix projection operation, σ denotes the Relu function and $\oplus$ denotes element-wise addition. Finally, our layered affinity matrix Z is computed as follows:

$$Z = Q \otimes M \tag{12}$$

where $\otimes$ denotes Hadamard product. By computing the affinity matrix and fusing features in each layer, we allow the model to integrate encoder semantics with a growing context.

To enable each localized model to focus more on the virtual class most relevant to its own data, we propose privatizing the parameters of the affinity query and the affinity matrix within each client rather than sharing them globally, as shown in 4. The affinity query represents localized semantics specialized to each client's unique data characteristics. The affinity key represents global semantics from the shared global knowledge. Thus, personalizing the affinity matrix enables the optimization of compatibility between local features and global semantics. Formally, the parameters of the affinity query and the affinity matrix are updated based on local data:

$$\theta_{i,q}^{t+1} = \theta_{i,q}^{t} - \alpha \nabla \mathcal{L}_i(\theta_i; \mathcal{D}_i) \tag{13}$$

$$\theta_{i,m}^{t+1} = \theta_{i,m}^{t} - \alpha \nabla \mathcal{L}_i(\theta_i; \mathcal{D}_i) \tag{14}$$

while the parameters of the affinity key are updated based on global data:

$$\theta_{i,k}^{t+1} = \theta_{i,k}^{t} - \frac{1}{N}\sum_{i=1}^{N} \nabla \mathcal{L}_i(\theta_i; \mathcal{D}_i) \tag{15}$$

where $\theta_{i,q}^{t+1}$, $\theta_{i,m}^{t+1}$ and $\theta_{i,k}^{t+1}$ denotes the affinity query, the affinity aggregation matrix and the affinity key for the i^{th} site, respectively. The α denotes the learning rate and $\nabla \mathcal{L}_i$ represents the local gradient. The subscript t and $t+1$ represent the t and $t+1$ communication round, respectively.

Hence, clients can emphasize compatible semantics by localizing the affinity query and matrix while acquiring common knowledge from the global context via the shared affinity key.

IV. Experiments

A. Datasets and Metrics

We evaluate our method on two real-world medical image datasets for segmentation tasks: retinal fundus (RIF) images for optic disc and cup segmentation and endoscopic (EndoPolyp) images for polyp segmentation.

- **Retinal fundus image dataset.** This dataset comprises retinal images from four clinical centers [32]–[34] and contains {101, 159, 400, 400} images distributed across four sites. The dataset includes three categories: retinal disc, retinal cup, and background.
- **Endoscopic polyp image dataset.** This dataset includes 2187 images gathered from four distinct sources [35]–[38] distributed among four clients with {1000, 380, 196, 612} images each. The dataset includes two categories: foreground polyp and background.

For both datasets, we allocate 70% for training, 10% for validation, and 20% for testing. The data is partitioned into patient-level subsets to enhance generalization. We preprocess images for both benchmark datasets by cropping them to 384×384 pixels and normalizing them to achieve zero mean and unit variance before feeding them into the network.

Metrics. Performance is evaluated on each local test set using metrics such as mean Intersection over Union (mIoU) and Hausdorff distance (HD). Higher mIoU values and lower HD values indicate superior segmentation results.

B. Implementation Details

We implemented our method in PyTorch using the Adam optimizer. The model was trained on both datasets with a learning rate of 0.001 and a batch size of 12. Training was conducted over 200 communication rounds, with each local client performing one epoch per round. The loss function balance factors λ_r and λ_v were both set to 1, ensuring equal weighting between real and virtual classes. The number of virtual classes was set to CN, where C represents the number of real classes (3 for the RIF, 2 for the EndoPolpy) and N represents the number of clients (4 for both datasets). All clients were trained in each round. U-Net [31] was used as the backbone for all methods, except for additional comparisons with FedDP [15].

Since FedDP is designed for the ViT architecture, we used the same architecture (PVTv2-b0 [39] with FPN [40]) to ensure a fair comparison. We added CN convolutional kernel weights as virtual classes in the final classification head. Our PAA design in ViT aligns with that in U-Net, inserted during the upsampling process. The ViT models were trained using the same hyperparameters as the U-Net in our experiments.

TABLE I: Comparison of personalized federated learning methods on retinal fundus images segmentation and endoscopic polyp images segmentation. The symbol † indicates that we use the same backbone network architecture of PVTv2 and FPN with FedDP. The best results are in **bold**. The experimental results demonstrate that our proposed approach achieves superior performance compared to existing methods in the majority of evaluations.

Method	Backbone	A	B	C	D	Avg.	A	B	C	D	Avg.
		mIoU (%) ↑					HD ($pix.$) ↓				
		Task1: Retinal Fundus Image Segmentation									
FedAvg	U-Net	67.93	64.35	76.43	80.83	72.39	35.12	47.93	24.33	13.18	30.14
FedRep		62.81	72.58	78.01	80.52	73.48	22.23	29.16	23.51	12.9	21.95
FedBABU		65.73	71.4	75.82	80.66	73.4	18.58	40.99	41.18	9.89	27.66
FedLC		70.49	71.34	77.73	80.91	75.11	17.4	23.4	19.29	8.78	17.21
IOP-FL		67.95	70.45	78.16	79.72	74.07	19.51	22.97	42.16	8.3	23.23
Ours		**71.91**	**72.82**	**79.42**	**81.98**	**76.53**	**15.34**	**17.93**	**14.24**	**7.89**	**13.85**
FedDP	PVTv2	70.45	77.04	78.57	80.11	76.54	18.63	13.36	14.51	9.1	13.9
Ours†		**73.04**	**81.47**	**79.78**	**81.26**	**78.89**	**16.59**	**11.66**	**12.94**	**8.8**	**12.5**
		Task2: Endoscopic Polyp Image Segmentation									
FedAvg	U-Net	72.83	52.11	67.02	57.96	62.48	4.45	29.4	21.66	17.72	18.31
FedRep		65.33	56.7	56.48	49.33	56.96	4.48	56.17	47.31	27.93	33.97
FedBABU		66.25	50.01	64.33	53.47	58.51	5.31	60.33	39.89	24.61	32.53
FedLC		72.62	62.06	67.34	58.6	65.15	4.09	30.01	15.23	21.93	17.81
IOP-FL		72.36	60.33	61.36	59.15	63.3	4.16	34.48	15.36	17.41	17.85
Ours		**74.17**	**64.1**	**70.24**	**60.18**	**67.17**	**3.79**	**28.67**	**14.68**	**16.33**	**15.87**
FedDP	PVTv2	73.85	67.02	68.6	60.36	67.48	4.35	**4.56**	8.08	24.28	10.31
Ours†		**76.38**	**68.42**	**72.18**	**61.85**	**69.71**	**4.17**	5.5	**5.69**	**17.32**	**8.17**

TABLE II: Ablation studies on the effectiveness of virtual class and personalized affinity aggregation.

VC	PAA	RIF		EndoPolyp	
		mIoU ↑	HD↓	mIoU ↑	HD↓
		72.39	30.14	62.48	18.31
✓		74.69	19.35	65.15	16.54
	✓	73.83	19.02	65.28	19.36
✓	✓	**76.53**	**13.85**	**67.17**	**15.87**

TABLE III: Ablation studies on the PAA module. We investigated the impact of parameter personalization in the PAA module with and without virtual classes.

VC	Personalized	RIF		EndoPolyp	
		mIoU↑	HD↓	mIoU↑	HD↓
		71.74	32.5	59.01	24.19
	✓	73.83	19.02	65.28	19.36
✓		71.47	23.06	62.75	21.21
✓	✓	**76.53**	**13.85**	**67.17**	**15.87**

C. Comparisons with State-of-the-arts

Compare methods. We compare our approach with the state-of-the-art PFL methods including FedRep [41] , FedBABU [14], FedLC [16] , IOP-FL [42] and FedDP [15]. We also use FedAvg [18] as the baseline PFL method for comparison. It merits note that FedDP employs vision transformers (ViT) architecture to achieve long-range dependency modeling. Therefore, we also modified our network to a ViT backbone to enable a fair comparison.

Evaluation on RIF dataset. For the retinal fundus optic disc and cup segmentation task, our approach achieved optimal performance at each site using a U-Net framework, improving mIoU by 1.42% and reducing HD by 3.36% on average, which is shown in I. Moreover, our improved method achieves remarkable results when compared to FedDP [15], which utilizes a Feature Pyramid Network (FPN) [40] with a Pyramid Vision Transformer backbone (PVTv2-b0) [39]. It shows that our virtual classes and personalized affinity aggregation consistently exhibit the capability to regularize localized representations and selectively aggregate global contextual knowledge across various backbones.

Evaluation on EndoPolyp dataset. For polyp segmentation, our method again achieves superior performance across different network architectures. This success is also attributed to our proposed methods, which effectively balance localization and globalization learning. Although our method did not achieve the best HD at site B when compared to FedDP, our mIoU score at this site was significantly higher, demonstrating the superiority of our approach. Additionally, the results in I indicate limited performance gains from some methods. This stems from their failure to account for personalization under semantically aligned local and global domains.

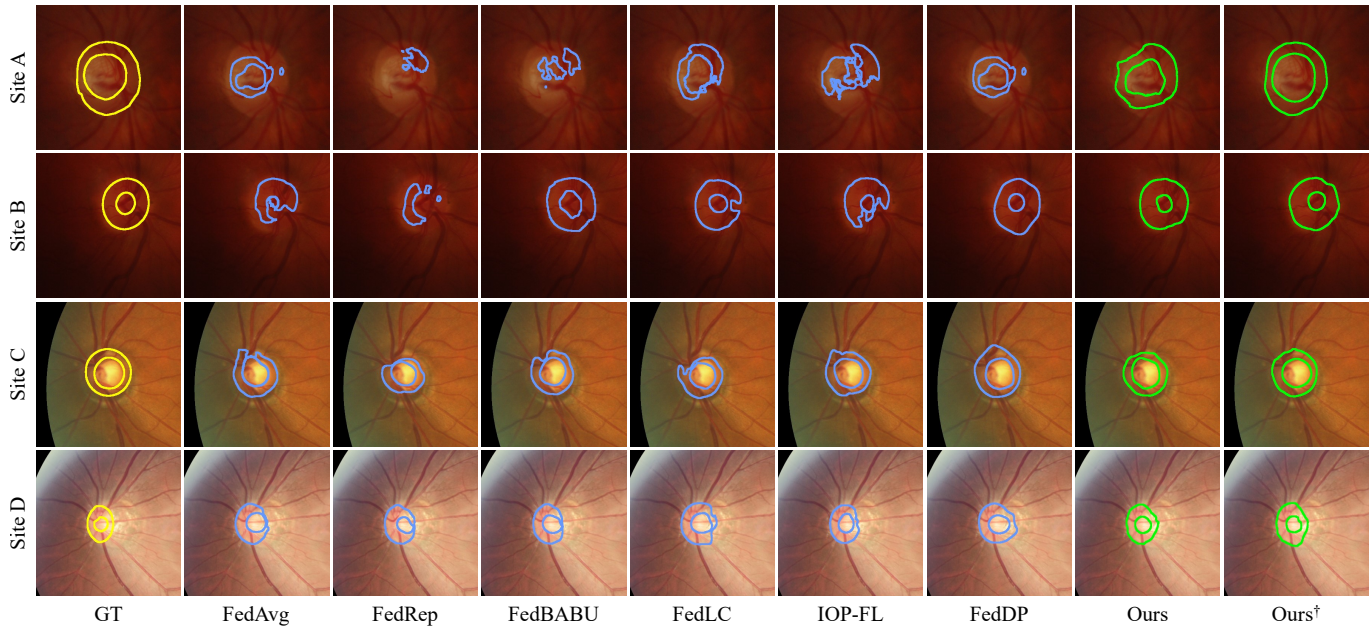

Fig. 5: Visual comparisons on the RIF dataset are provided between our approach (green) and current state-of-the-art PFL methods (blue). Each row depicts a randomly selected sample from a unique particular site.

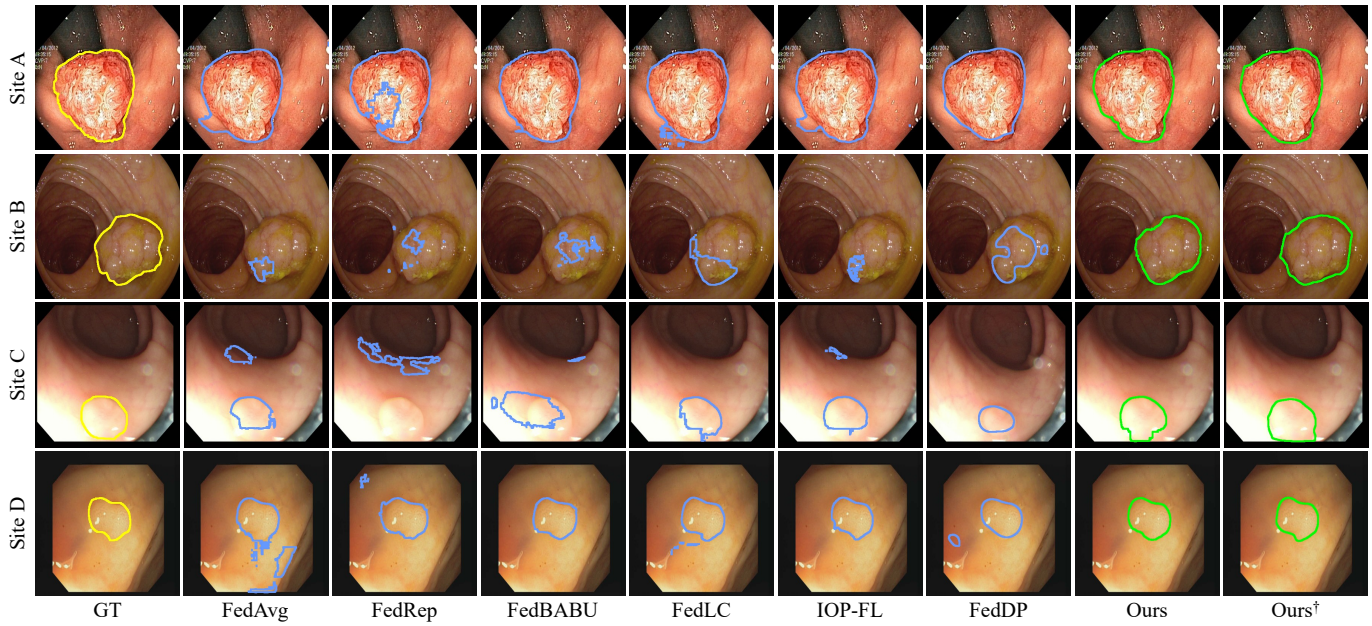

Fig. 6: Visual comparisons on the EndoPolyp dataset are provided between our approach (green) and current state-of-the-art PFL methods (blue). Each row depicts a randomly selected sample from a unique particular site.

Visualized results analysis. To better exhibit the superiority of our approach and performance improvements conferred, we provide visualizations of segmentation results on randomly selected samples from each site. The retinal fundus optic disc and cup segmentation results are shown in Fig. 5. Our approach exhibits significant segmentation performance improvements on smaller datasets from sites A and B. Especially in scenarios where the distinction between foreground and background is ambiguous, our method achieves superior delineation. This is attributed to our method's ability to preserve personalized localization semantics without distortion while acquiring global knowledge. Additionally, our method achieves tighter contour adhesion on larger datasets from Sites C and D and captures finer-grained anatomical intricacies. This is attributed to our method's capability to extract more localized details from more data samples while retaining the transfer of generalized knowledge. In addition, we provide visualizations of the segmentation results for the polyp dataset. As depicted in the Fig. 6, our method demonstrates effective segmentation performance for both large and small polyps,

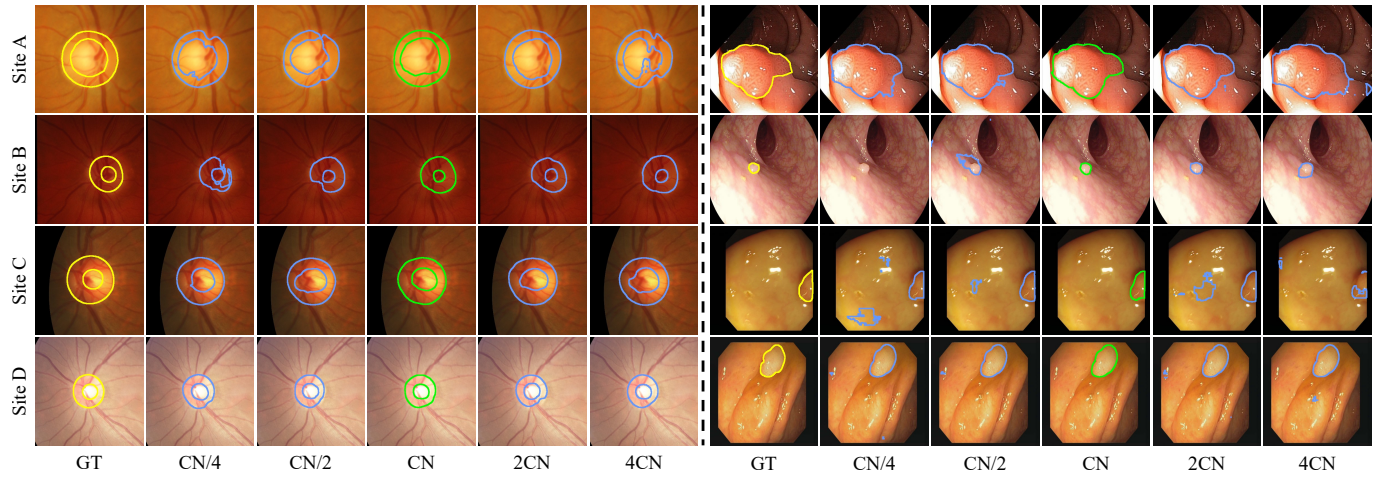

Fig. 7: Visual comparison of segmentation results on different virtual classes. We randomly select samples from each site on RIF (left) and EndoPolyp (right) datasets. The green represents the best results and the blue for the rest of the settings.

when dealing with datasets that present substantial variations across sites.

D. Ablation Studies

In this section, we aim to validate the core insights of our approach by adding or removing proposed components, *i.e.*, virtual classes (VC), and personalized affinity aggregation (PAA). Furthermore, we discuss the role of VC and PAA in boosting segmentation efficacy.

Effectiveness of each component. As II presents the results for each module. Incorporating virtual classes significantly improves mIoU scores and reduces HD on both datasets, demonstrating the effectiveness of learning heterogeneous site data of the same class with multiple virtual classes. Adding personalized affinity aggregation further increases mIoU and decreases HD, validating its role in refining localization by filtering inconsistent global knowledge. However, we notice that solely incorporating PAA leads to insignificant improvements, especially on EndoPolyp datasets with severe domain shifts. In contrast, adding VC realizes consistent performance gains. This implies VC has a more pivotal role in effective personalization. Without concrete virtual classes to guide affinity aggregation, the localized affinity computations lack target concepts to transfer. The virtual classes establish more meaningful embeddings for the real classes, thereby facilitating the effectiveness of PAA module.

T-SNE visualization. To intuitively demonstrate the performance gains from our virtual classes, we provide a t-SNE visual comparison. As shown in Fig. 8 (a) and (b), initial feature (a) exhibit semantic ambiguity in the feature space without other techniques. In contrast, our method(b) achieves more coherent clustering under the influence of virtual classes.

Optimal virtual classes number. A key hyperparameter in our approach is selecting the number of virtual classes. Intuitively, we totally allocate CN virtual classes for each client class to allow personalized representations. As shown in Fig. 8 (c) and (d), we explore using {CN/4, CN/2, CN, 2CN, 4CN} virtual classes and CN virtual classes to achieve the optimal balance between localization and globalization. Using fewer virtual classes constrains adaptability, while excessive classes fragment global semantics.

Additionally, we visualize segmentation outcomes using varying numbers of virtual classes to provide insights, which is shown in Fig. 7. With too few virtual classes, segmentation boundaries become less smooth and coherent. This is because fewer virtual classes cannot capture necessary shape and texture variations within the same class in heterogeneous data. Conversely, an overabundance of virtual classes also degrades segmentation quality and consistency. In such situations, features of the same class become overly detailed, hindering the extraction of generalized information from the global context.

Effectiveness of PAA. As depicted in Fig. 8 (e) and (f), we systematically analyze the impact of sharing each affinity parameter while localizing the others, including the affinity query (Q), the affinity key (K), and the affinity matrix (M). Sharing the encoder affinity key with localized decoder affinity query and the affinity matrix achieves optimal segmentation performance. The shared encoder affinity key ensures consistent semantic targets across all clients. The localized decoder affinity query allows each client to extract only the most relevant global representations tailored to its domain. The personalized affinity matrix enhances flexibility for sites to locally adapt feature fusion to their specific needs.

In contrast, sharing the decoder affinity query may result in the model overfitting to sites with an abundance of samples, compromising its ability to learn fine-grained details effectively. Similarly, sharing the affinity matrix alone may lead to a loss of flexibility in adapting feature fusion locally, impeding the model from capturing specific local details.

Furthermore, we analyze the impact of personalization within the PAA module, which is shown in Tab. III. The results highlight significant improvements achieved through the customization of PAA parameters. Non-personalized affinity aggregation faces challenges in capturing dataset nuances, hindering the on-site personalization capability brought by virtual classes.

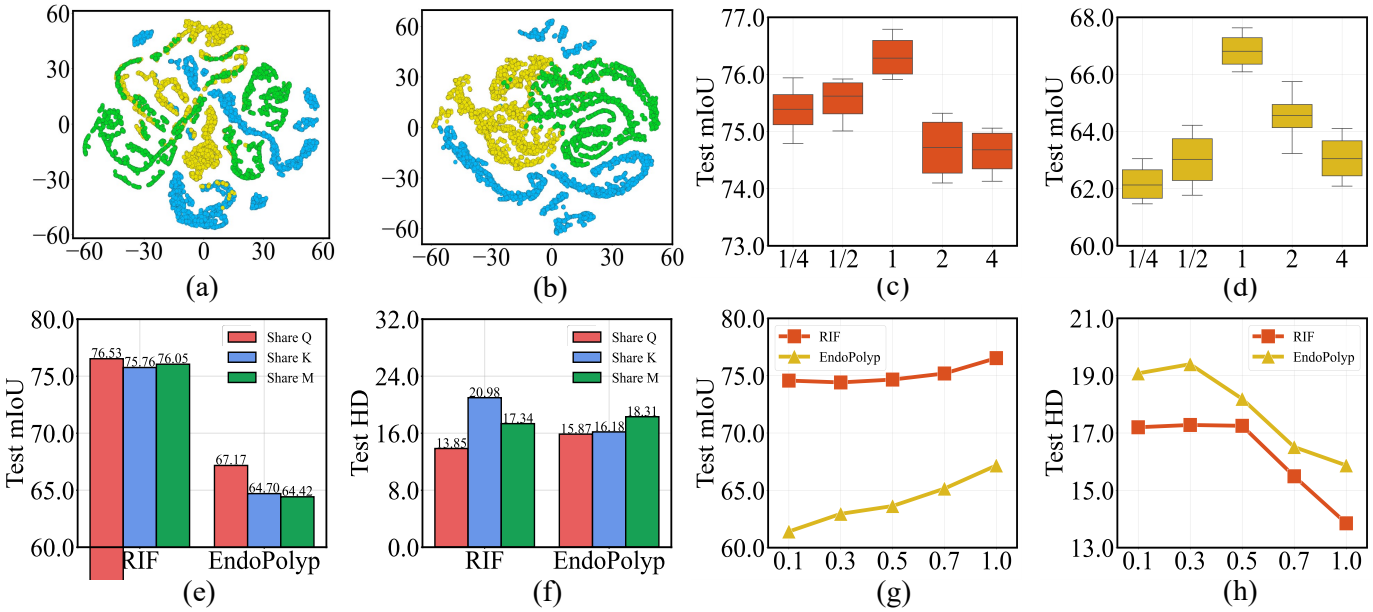

Fig. 8: Analysis of our method. (a)-(b) T-SNE visualization of feature embeddings from different sites before (a) and after training (b). Different colors represent different categories. (c) Test mIoU variation with the number of virtual classes on the RIF dataset and (d) on the EndoPolyp dataset. We overlooked the unit (CN) on the x-axis. (e)-(f) The impact of sharing different component parameters. (g)-(h) The test mIoU and HD vary with the change in the weight of virtual class loss. The weight of real class loss remains at 1.

TABLE IV: Comparison with state-of-the-art approaches in terms of efficiency and privacy

Extra	FedAvg	FedLC	FedDP	Ours
Communication	–	✓	✓	–
Training	–	✓	✓	–

Effectiveness of virtual classes loss. As shown in Fig. 8 (g) and (h), we studied the impact of using different weights for virtual classes. We found that setting both weights to 1 achieves the best performance. Using a lower weight for the virtual classes potentially hinders the model's ability to build personalized representations that capture nuances for each real class. Meanwhile, maintaining an equal weight of 1 for the real classes preserves full supervision signals to discriminate anatomical structures fundamental for segmentation.

E. More Discussion

Privacy. To ensure a fair comparison, we employed the same training protocol (one epoch per round) for all methods. This standardization allows for a direct comparison of the effectiveness of different approaches. Additionally, our method utilizes a widely used communication protocol that only involves essential parameter exchanges between clients and the server. Consequently, we reduce privacy risks compared to state-of-the-art methods such as FedDP [15] and FedLC [16], which involve additional client-to-client communications, as shown in Tab. IV.

Efficiency analysis. The latest personalized medical image segmentation approaches harness inter-site inconsistencies for learning. FedLC [16] fine-tunes using other sites' segmentations, while FedDP [15] detects inconsistencies by amassing parameters from all clients. Both entail additional N-1 training per local iteration where N is the number of sites, hampering efficiency at scale. In contrast, our method neither necessitates parameter collection nor external fine-tuning. We simply append convolution kernels in the segmentation head, readily expanding to myriad clients while guaranteeing high computational efficiency.

Limitation. Our approach still has limitations, particularly in balancing scene complexity with a limited number of virtual classes. Introducing too many virtual classes can lead to overfitting at individual sites, which may fragment the comprehensive semantic representation of classes. Conversely, having too few virtual classes can result in insufficient representation of complexity scene.

V. Conclusion

In this work, we introduce a novel strategy for personalized federated medical image segmentation, which aims to mitigate data heterogeneity across clients. Our approach utilizes immutable virtual classes for global alignment with personalized allocation. Additionally, we develop a personalized affinity aggregation module, maintaining on-site recognition of virtual classes while benefiting from global knowledge. Extensive experiments on two medical datasets validate the effectiveness of our techniques.

REFERENCES

[1] J. Wicaksana, Z. Yan, D. Zhang, X. Huang, H. Wu, X. Yang, and K.-T. Cheng, "Fedmix: Mixed supervised federated learning for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 1955–1968, 2022.

[2] B. Lei, Y. Zhu, E. Liang, P. Yang, S. Chen, H. Hu, H. Xie, Z. Wei, F. Hao, X. Song *et al.*, "Federated domain adaptation via transformer for multi-site alzheimer's disease diagnosis," *IEEE Transactions on Medical Imaging*, 2023.

[3] Z. Gürler and I. Rekik, "Federated brain graph evolution prediction using decentralized connectivity datasets with temporally-varying acquisitions," *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, 2022.

[4] X. Xu, H. H. Deng, J. Gateno, and P. Yan, "Federated multi-organ segmentation with inconsistent labels," *IEEE transactions on medical imaging*, vol. 42, no. 10, pp. 2948–2960, 2023.

[5] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[6] Q. Wu, K. He, and X. Chen, "Personalized federated learning for intelligent iot applications: A cloud-edge based framework," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 35–44, 2020.

[7] Z. Chen, C. Yang, M. Zhu, Z. Peng, and Y. Yuan, "Personalized retrogress-resilient federated learning toward imbalanced medical data," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3663–3674, 2022.

[8] V. Kulkarni, M. Kulkarni, and A. Pant, "Survey of personalization techniques for federated learning," in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE, 2020, pp. 794–797.

[9] J. Xu, X. Tong, and S.-L. Huang, "Personalized federated learning with feature alignment and classifier collaboration," *arXiv preprint arXiv:2306.11867*, 2023.

[10] M. Chen, M. Jiang, Q. Dou, Z. Wang, and X. Li, "Fedsoup: Improving generalization and personalization in federated learning via selective model interpolation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 318–328.

[11] O. Marfoq, G. Neglia, A. Bellet, L. Kameni, and R. Vidal, "Federated multi-task learning under a mixture of distributions," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 434–15 447, 2021.

[12] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.

[13] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," *arXiv preprint arXiv:2102.07623*, 2021.

[14] J. Oh, S. Kim, and S.-Y. Yun, "Fedbabu: Towards enhanced representation for federated image classification," *arXiv preprint arXiv:2106.06042*, 2021.

[15] J. Wang, Y. Jin, D. Stoyanov, and L. Wang, "Feddp: Dual personalization in federated medical image segmentation," *IEEE Transactions on Medical Imaging*, 2023.

[16] J. Wang, Y. Jin, and L. Wang, "Personalizing federated medical image segmentation via local calibration," in *European Conference on Computer Vision*. Springer, 2022, pp. 456–472.

[17] H. Kassem, D. Alapatt, P. Mascagni, A. Karargyris, and N. Padoy, "Federated cycling (fedcy): Semi-supervised federated learning of surgical phases," *IEEE transactions on medical imaging*, vol. 42, no. 7, pp. 1920–1931, 2022.

[18] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[19] X. Gong, L. Song, R. Vedula, A. Sharma, M. Zheng, B. Planche, A. Innanje, T. Chen, J. Yuan, D. Doermann *et al.*, "Federated learning with privacy-preserving ensemble attention distillation," *IEEE transactions on medical imaging*, vol. 42, no. 7, pp. 2057–2067, 2022.

[20] D. Chen, J. Hu, V. J. Tan, X. Wei, and E. Wu, "Elastic aggregation for federated optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 187–12 197.

[21] H. Wu, B. Zhang, C. Chen, and J. Qin, "Federated semi-supervised medical image segmentation via prototype-based pseudo-labeling and contrastive learning," *IEEE Transactions on Medical Imaging*, 2023.

[22] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 713–10 722.

[23] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, and H. Guan, "Fedala: Adaptive local aggregation for personalized federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, 2023, pp. 11 237–11 244.

[24] J. Zhang, S. Guo, X. Ma, H. Wang, W. Xu, and F. Wu, "Parameterized knowledge transfer for personalized federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 092–10 104, 2021.

[25] Z. Li, X. Shang, R. He, T. Lin, and C. Wu, "No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier," *arXiv preprint arXiv:2303.10058*, 2023.

[26] Y. Wang, H. Xu, W. Ali, M. Li, X. Zhou, and J. Shao, "Fedftha: a fine-tuning and head aggregation method in federated learning," *IEEE Internet of Things Journal*, 2023.

[27] B. Chen, W. Deng, and H. Shen, "Virtual class enhanced discriminative embedding learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[28] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, L. Ma, S. Pu, and D.-C. Zhan, "Forward compatible few-shot class-incremental learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9046–9056.

[29] Z. Song, Y. Zhao, Y. Shi, P. Peng, L. Yuan, and Y. Tian, "Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 183–24 192.

[30] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[32] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "Rim-one: An open retinal image database for optic nerve evaluation," in *2011 24th international symposium on computer-based medical systems (CBMS)*. IEEE, 2011, pp. 1–6.

[33] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish *et al.*, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.

[34] J. I. Orlando, H. Fu, J. B. Breda, K. Van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee *et al.*, "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical image analysis*, vol. 59, p. 101570, 2020.

[35] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[36] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International journal of computer assisted radiology and surgery*, vol. 9, pp. 283–293, 2014.

[37] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[38] 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," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*. Springer, 2020, pp. 451–462.

[39] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[40] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[41] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International conference on machine learning*. PMLR, 2021, pp. 2089–2099.

[42] M. Jiang, H. Yang, C. Cheng, and Q. Dou, "Iop-fl: Inside-outside personalization for federated medical image segmentation," *IEEE Transactions on Medical Imaging*, 2023.