FedDP: Dual Personalization in Federated Medical Image Segmentation

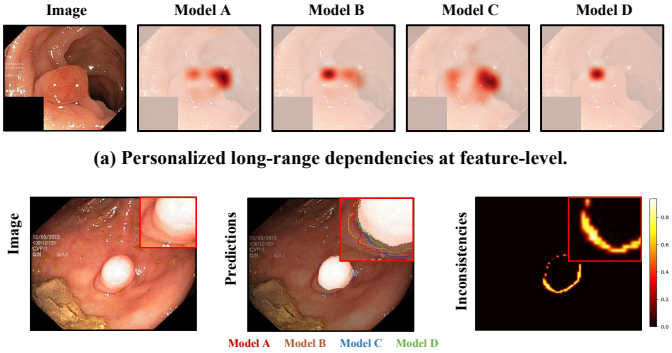
Jiacheng Wang, *Student Member, IEEE*, Yueming Jin, *Member, IEEE*, Danail Stoyanov, *Senior Member, IEEE*, Liansheng Wang, *Member, IEEE*

***Abstract*— Personalized federated learning (PFL) addresses the data heterogeneity challenge faced by general federated learning (GFL). Rather than learning a single global model, with PFL a collection of models are adapted to the unique feature distribution of each site. However, current PFL methods rarely consider self-attention networks which can handle data heterogeneity by long-range dependency modeling and they do not utilize prediction inconsistencies in local models as an indicator of site uniqueness. In this paper, we propose *FedDP*, a novel *fed*erated learning scheme with *d*ual *p*ersonalization, which improves model personalization from both feature and prediction aspects to boost image segmentation results. We leverage long-range dependencies by designing a *local query* (LQ) that decouples the query embedding layer out of each local model, whose parameters are trained privately to better adapt to the respective feature distribution of the site. We then propose *inconsistency-guided calibration* (IGC), which exploits the inter-site prediction inconsistencies to accommodate the model learning concentration. By encouraging a model to penalize pixels with larger inconsistencies, we better tailor prediction-level patterns to each local site. Experimentally, we compare FedDP with the state-of-the-art PFL methods on two popular medical image segmentation tasks with different modalities, where our results consistently outperform others on both tasks. Our code and models will be available at `https://github.com/jcwang123/PFL-Seg-Trans`.**

***Index Terms*— Personalized federated learning; Medical image segmentation; Self-attention mechanism**

I. Introduction

Federated learning (FL) is an important topic in medical image computing enabling model training while adhering to local data protection rules [1], [2]. Recent studies have achieved remarkable success in various lesion segmentation tasks, e.g., brain tumors [3], [4], polyp lesions [5] and multiple organ segmentation tasks [6]–[8]. General federated learning (GFL) allows multiple participating data sites to train a single global model without centralizing their data. Knowledge aggregation is achieved through averaging all local parameters on a cloud server [9]. However, a single global model can be sub-optimal in handling heterogeneous data distributions and may result in performance bias among different sites due to individual distribution variation frequently occurring in practice [10]. A single global model can therefore struggle to generate accurate or optimal predictions for each site [11]. It is also possible that averaging local parameters in one cloud server can be susceptible to cyber attacks and training time attacks from malicious data sites [12], [13].



(a) Personalized long-range dependencies at feature-level.

(b) Exploring inter-site inconsistencies at prediction-level.

Fig. 1. Illustration of the intuition behind our proposed method **FedDP**. We use the learned different local self-attention models (Model A to D) to infer the same image sample and visualize the (a) attention maps for the center pixel and (b) predicted boundaries. **FedDP** performs dual personalization focusing on these two aspects, (a) long-range dependency modeling (feature-level) and (b) inter-site inconsistencies (prediction-level) to improve local segmentation. Note that we use local models to predict the same sample only for illustration. Data in one site is not transferred to others in the real federation training process.

Personalized federated learning (PFL) proposes to learn multiple local models rather than a single model, and each local model aims to adapt to a unique feature distribution or a tailored prediction pattern for each site [11]. PFL proposes to decouple partial model parameters out of the inter-site communication and only learn them locally. For example, some methods decouple parameters in the feature extractor to fit feature distributions, such as the Batch Normalization (BN) layers [14] and the high-frequency components of convolutional parameters [15]. Most existing methods propose to learn personalized prediction layers for each local model, e.g., the last full-connection layer in the network [5], [16], [17]. Another stream of methods treats the model as a whole and utilizes a weight decay strategy to save historical local models [18], [19]. Each local model can be better adapted to

Jiacheng Wang and Liansheng Wang are with the Department of Computer Science at School of Informatics, Xiamen University, Xiamen 361005, China. (e-mail:jiachengw@stu.xmu.edu.cn, lswang@xmu.edu.cn)

Yueming Jin and Danail Stoyanov are with the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Computer Science, University College London, UK. (e-mail: {yueming.jin, danail.stoyanov}@ucl.ac.uk)

Corresponding author: Liansheng Wang (lswang@xmu.edu.cn)

its corresponding data distribution, by learning more intra-site knowledge from historical models.

Despite recent progress, existing PFL approaches for image segmentation have limited feature personalization regarding modeling long-range dependencies, since they are designed based on Convolutional Neural Networks (CNNs) with limited convolutional kernel size. However, modeling long-range dependencies has been widely demonstrated its efficacy in various vision tasks which can aggregate the contexts from the whole image space for better representation learning. This can potentially be addressed by **self-attention networks**, also known as vision transformers [20]–[24], which can perform feature extraction via long-range dependency modeling with strong robustness proven to handle the distribution shifts in self-supervised learning [25], multi-modality learning [26], and general federated learning [27]. For architectures based on long-range dependencies, personalizing their long-range modeling plays a core role in the local feature distribution adaption, however, it is still under-explored for medical segmentation tasks. Furthermore, the presence of diverse imaging protocols with varying settings across different clinical sites can result in variations in the boundaries of the same target. Moreover, different raters are prone to annotating the target boundary with noticeable discrepancies [28], resulting in inconsistent segmentation labels. These factors contribute to distinct optimization paths in individual local models and, consequently, significantly undermine performance when aggregating the parameters [11]. Such inconsistency across different sites tends to highlight the respective uniqueness of each site. Therefore, leveraging such inter-site inconsistency, for example, the inconsistent results generated by different local models, can benefit the perception of the tailored prediction pattern in each local site. However, existing methods rarely consider utilizing this information when personalizing predictions, and the most relative attempt [5] introduces extra computational costs during the inference.

In this paper, we propose a novel personalized federated learning method, named **FedDP**, to comprehensively address the **d**ual **p**ersonalization to boost the local segmentation performance. (i) For the feature-level personalization, **FedDP** proposes the local query (LQ) which decouples query/key embedding layers of the self-attention networks, and saves the query portion as the local part without sharing it with other sites during the model learning. As query embeddings generally represent each pixel's specialized features while key embeddings denote supportive features from other pixels, locally training query embedding layers helps each site with exploring its special long-range clues. (ii) For the prediction-level personalization, **FedDP** proposes the inconsistency-guided calibration (IGC) to leverage the inter-site prediction inconsistencies. Concretely, we compute pixel-level inconsistencies of each sample by gathering all local models to predict the segmentation maps and then calculating disagreements of maps. We encourage the model attending to pixels with large inconsistencies, since such pixels inherently highlight each site's special prediction patterns. It is achieved by designing a new loss to feed more supervision on inconsistent pixels, and no extra computation costs are brought during the inference.

Our method is evaluated on two medical image segmentation tasks with different modalities, including polyp segmentation from endoscopic images and optic disc/cup segmentation from retinal fundus images. We compare our method with several GFL methods and the latest PFL methods, where our method has consistently achieved superior performance.

II. Related Work

A. Personalized Federated Learning

Model personalization in federated learning aims to learn unique local models specialized to each distinct feature distribution and prediction custom, through which the local accuracy can be improved and different solution demands are satisfied [11]. There are a variety of methods to achieve this goal by local finetuning [29]–[31], meta-learning [32], knowledge distillation [33], weight decay [18], [19], and model decoupling [5], [14]–[17]. The simplest strategy, local finetuning, tunes each local model's parameters by training them for a few epochs on the local dataset after the federated learning process, obtaining large improvements in the local accuracy especially of the sites with imperfect federation accuracy. According to which parameters it tunes, the methods can achieve different personalization goals, i.e., tuning prediction layers denotes the prediction-level personalization. The methods using meta-learning, knowledge distillation, and weight decay, additionally build a set of local models that are entirely different from the global modal and transmit the global knowledge into the local models, thus significantly reducing the negative effects of the global bias. Another group proposes not additionally training the local models, but decoupling partial parameters to be saved locally [14]–[17] and globally sharing the rest of parameters. The local parameters are used to tune the features by personalizing normalization layers [14] and convolutional layers [15], and to tune the predictions by personalizing the prediction layers [16], [17]. However, all these methods are designed for convolutional networks and lack the ability to personalize the long-range dependencies and predictions for self-attention-based networks. In addition, the prediction personalization is not well handled as they lack the exploration of inter-site prediction inconsistencies. More recently, the CNN-based work [5] proposes to explore the inter-site inconsistencies by transmitting all local prediction layers into each local site at each federation round, which increases the transmission costs. What's more, it uses inconsistencies through an attention module which also increases the computational complexity. In comparison, FedDP requires one-time transmission and utilizes the inconsistency to re-weight the supervised loss which is efficient in terms of the transmission and computation.

B. Self-Attention Networks

Self-attention architectures are firstly proposed for the sequence-to-sequence machine translation [34] and subsequently extended to other Natural Language Processing (NLP) tasks. Recently, they have been broadly applied in the image processing field to model the long-range dependencies in the whole view of an image, which are also known as

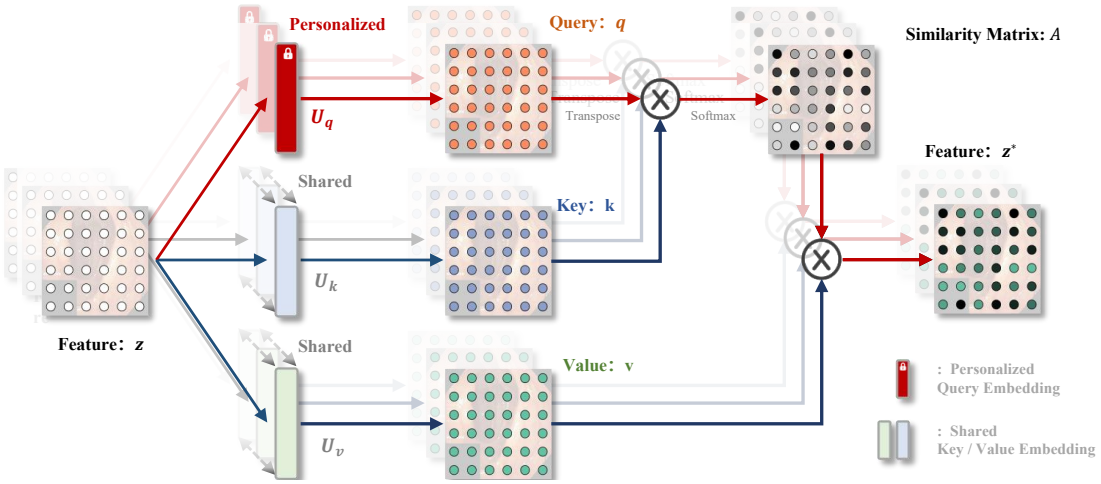


Fig. 2. Visualization of the **feature-level personalization** by decoupling the query embedding layers $\mathbf{U_q}$ (red) out of the self-attention networks to be trained locally. Instead, the key embedding $\mathbf{U_k}$ and the value embedding layers $\mathbf{U_v}$ (blue and green) are shared in all local sites. The unique long-range dependency modeling paradigm can be constructed in each local site, contributing to a better feature distribution adaptation.

vision transformers. For example, the first vision transformer, ViT [35], achieves the best performance on ImageNet classification by splitting an image into patches and applying self-attention to catch the global contexts. The following work improves the architecture by enhancing the multi-scale representations and local context modeling ability [36], [37] and is applied to image denoising [38], segmentation [39], detection [40], [41]. Medical image segmentation also benefits from the advancement of self-attention networks, i.e., 3D brain tumor segmentation [20], skin lesion segmentation [21], [24], polyp segmentation [22], and other areas [23], [42]. It proves that self-attention networks can handle well targets with large shape and size variance and have better robustness on unseen data compared to convolutional networks. However, these networks are all trained by only using the data within one local site, not considering training the models using the multi-institution data with patient privacy protection. One latest work [27] empirically applies self-attention networks to federated classification tasks and finds that their performance in the heterogeneous setting has a significantly larger increase in the convergence speed and test accuracy compared to the convolutional networks. Yet, this work addresses the heterogeneity using general FL algorithms and has not implemented them on the segmentation tasks which are more challenging with pixel-level representations. Hence, we fill this gap and propose to personalize the long-range dependency modeling to enhance the local accuracy when applying self-attention networks to PFL segmentation tasks.

III. Method

To personalize federated medical image segmentation, we propose a novel PFL framework **FedDP**, accomplishing the personalization at dual feature- and prediction- levels. The overall learning framework is described in Section III-A. The feature-level personalization and prediction-level personalization are then introduced in Section III-B and III-C, respectively.

A. Overall Pipeline

We start with a brief background explanation of the general federated learning (GFL) and existing personalized federated learning (PFL) solutions, and then describe the pipeline of our federation **FedDP** with dual personalization.

1) Backgrounds: Assumed that there are K local sites with their unique distributions $\{\mathcal{D}_k\}_{k=1}^K$ and the joint image and label space is denoted as $\{(\mathcal{X}_k, \mathcal{Y}_k)\}_{k=1}^K$. For the k-th site, a set of N_k samples $\{(x_{i,k}, y_{i,k})\}_{i=1}^{N_k}$ establish the non-iid distribution $\mathcal{D}_k$, where $(x_{i,k}, y_{i,k}) \in \mathcal{D}_k$. The most typical general FL framework, i.e., FedAVG [9], aims to learn a single global model w to suit all distributions by first collecting the gradients from all local sites and then doing the average, as

$$w^{t+1} = w^t - \eta \mathbb{E}_{k\in[K]} \mathbb{E}_{(x_{i,k},y_{i,k})\in\mathcal{D}_k} \nabla \mathcal{L}(f_w(x_{i,k}), y_{i,k}). \quad (1)$$

Here, f_w is the model function parameterized by w; $\mathcal{L}$ is the loss function; η is the learning rate and t denotes the communication round.

As a single model can not satisfy the distinct solutions owing to the distribution and prediction difference, general FL methods usually suffer from poor local accuracy when the inter-site drift is large. Instead, PFL methods propose to learn the respective local models $\{v_k\}_{k=1}^K$ for all the local sites, where each local model v_k aims to suit the k-th site's special feature distribution and prediction custom.

2) Federation with Dual Personalization: **FedDP** proposes dual personalization with local query (LQ) and inconsistency-guided calibration (IGC), to respectively enhance FL from two perspectives, i.e., long-range dependency modeling and inter-site prediction inconsistency exploration. The overall pipeline is shown in Algorithm 1. It contains two learning stages that are the dependency personalization stage and the prediction calibration stage, where LQ and IGC are respectively performed. During the stage of dependency personalization, all local models undergo parallel training on their respective local devices. Subsequently, the globally shared parameters are transmitted to a cloud server, where their averaged values are computed. These averaged parameters are then sent back to the local sites and combined with the locally personalized

Algorithm 1: Federated learning with dual personalization (**FedDP**).

Input: Local datasets $\{(\mathcal{X}_k, \mathcal{Y}_k)\}_{k=1}^K$, Number of communication rounds T.

Output: Personalized local models $\{v_k^*\}_{k=1}^K$.

Initialize *query* parameters for each model $\{\rho_k^0\}_{k=1}^K$;
Initialize the rest parameters shared by all models γ^0;
/* Dependency Personalization */
for $t = 0, 1, ..., T-1$ **do**
  for $k = 1, 2, ..., K$ **do**
    Send γ^t to each site ;
    $\hat{\gamma}_k^{t+1}, \rho_k^{t+1} \leftarrow \mathbf{LQ}((\mathcal{X}_k, \mathcal{Y}_k); \gamma^t, \rho_k^t)$;
  end
  $\gamma^{t+1} = \frac{1}{K}\sum_{j=1}^K \hat{\gamma}_j^{t+1}$;
end
/* Merge Parameters */
for $k = 1, 2, ..., K$ **do**
  $\hat{v}_k := \gamma^T \cup \rho_k^T$;
end
/* Prediction Calibration */
for $k = 1, 2, ..., K$ **do**
  Send $\{\hat{v}_j\}_{j=1}^K$ to each site;
  $v_k^* = \mathbf{IGC}((\mathcal{X}_k, \mathcal{Y}_k); \hat{v}_k, \{\hat{v}_j\}_{j=1}^K)$;
end

parameters, resulting in a set of preliminary local models., $\{\hat{v}_k\}_{k=1}^K$. At the second stage, each local site collects all the local models through one-time communication, and computes the inconsistencies among them on the each local device. Each model's parameters are then tuned under the guidance of inter-site inconsistencies, resulting in the finely calibrated models $\{v_k^*\}_{k=1}^K$ eventually.

B. Long-range Dependency Personalization

In the group of PFL methods, decoupling partial parameters to save locally is helpful to make the models adaptive to several unique distributions. There are several alternative types of layers to be decoupled, e.g., Batch Normalization layers [14] or Prediction layers [16], [17], regarding the feature personalization or prediction personalization. However, these solutions have not leveraged the great robustness of self-attention-based networks to enhance PFL. As their superior performance is heavily attributed to the ability of long-range dependency modeling, how to personalize long-range dependency to refine each site's feature distribution is crucial when developing a self-attention-based PFL framework. To fill this gap, **FedDP** proposes the local query embedding where each local self-attention model obtains its personalized query embedding custom and all local models share the key embedding standards.

Most self-attention models exploit the long-range dependency through the multi-head self-attention mechanism, based on the **qkv** calculation. Given an input image x, it is divided into several square patches rigidly and each patch is encoded by a Multi-Layer-Perception (MLP) module. The obtained patch embedding is concatenated with a positional embedding and flattened to form the sequential embeddings, which are then augmented through the cascaded self-attention blocks. Assuming that $\mathbf{z}$ denotes the input embedding of one self-attention block with the length of L, the **qkv** calculator firstly outputs the *query* (**q**), *key* (**k**), *value* (**v**), as $[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv}$, where $\mathbf{U}_{qkv}$ is a full-connection layer, and $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^L$. After that, each element (a.k.a. each pixel embedding) in query $\mathbf{q}_l$ is then compared to all the elements in key in the long-range view by computing the similarity matrix, as

$$\mathbf{A}_l = softmax(\mathbf{q}_l \mathbf{k}^\top). \tag{2}$$

Similarity weight in matrix $\mathbf{A}_l \in \mathbb{R}^L$ is multiplied with value $\mathbf{v}$ in element wise and all weighted contexts are then added as

$$\mathbf{z}_l^* = \mathbf{A}_l \mathbf{v}. \tag{3}$$

As shown in the equations above, for the l-th element $\mathbf{z}_l$, its enhancement $\mathbf{z}_l^*$ captures the long-range contexts majorly through the calculation of the similarity matrix $\mathbf{A}_l$. Hence, personalizing the matrix calculation process for each local site leads to personalized long-range dependencies. To achieve this goal, two choices are naturally considered, i.e., personalizing the query embedding or the key embedding. As shown in Eq. 2, the similarity matrix of the l-th element is calculated between the l-th query embedding and all key embeddings. To this end, the query embedding can represent the specialized semantic cues of each pixel, while the key embedding represents the supportive information from all pixels that is used to augment the current one.

Therefore, we propose the **local query** that each site preserves its personalized query embedding custom, while all sites share the key and the value embedding standards. Specifically, let v denote the entire parameter set of a transformer-based segmentation network, i.e., a feature pyramid network (FPN) [43] with the backbone of a pyramid vision transformer (PVT-b0) [44]. Let ρ denote the parameters of query embedding layers $\mathbf{U}_q$ (local) and γ denote the rest parameters (global), i.e., $v \rightarrow \rho \cup \gamma$ and $\rho \cap \gamma = \varnothing$. As shown in Fig. 2, local parameters are updated only by the local gradients as

$$\rho_k^{t+1} = \rho_k^t - \eta \mathbb{E}_{(x_{i,k}, y_{i,k}) \in \mathcal{D}_k} \nabla \mathcal{L}(f_{v_k}(x_{i,k}), y_{i,k}), \tag{4}$$

and the global parameters are optimized through the global gradients as

$$\gamma_k^{t+1} = \gamma_k^t - \eta \mathbb{E}_{k \in [K]} \mathbb{E}_{(x_{i,k}, y_{i,k}) \in \mathcal{D}_k} \nabla \mathcal{L}(f_{v_k}(x_{i,k}), y_{i,k}). \tag{5}$$

After training, each local site obtains its intermediate personalized parameters $\hat{v}_k$.

C. Calibrating Prediction with Inter-site Inconsistencies

Exploring the inconsistency knowledge among multiple annotations from different clinical experts or raters for an image sample has been proven to be able to enhance representation learning and improve segmentation results [28]. In the federated learning paradigm, clinical raters from different centers generally have different labeling standards, therefore, the trained local models can inherently generate inconsistent

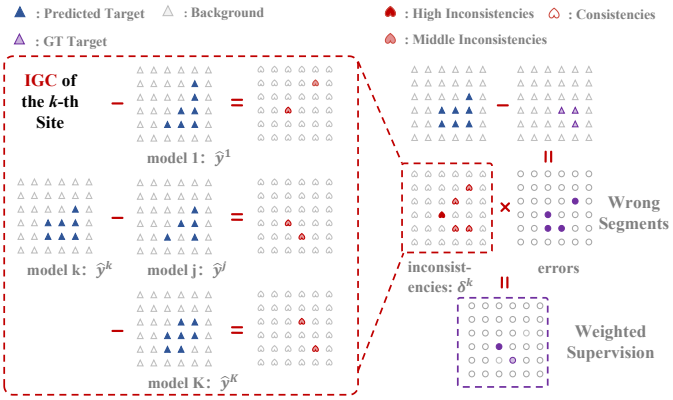


Fig. 3. Visualization of the proposed inconsistency-guided calibration (IGC) in the k-th site. It explores the inter-site prediction inconsistencies (red) and enlarges the supervision of pixels with large inconsistencies. Note that the red heart denotes the pixels with large inconsistencies and the white heart denotes the pixels with no inconsistencies. The errors are calculated through the entropy loss in our implementation and we simplify the process using the minus in this figure.

predictions with rich information, which can strengthen the regions holding more uniqueness in each local site. However, most existing PFL methods overlook the potential of this inter-site knowledge. We propose an inconsistency-guided calibration (IGC), which perceives the inter-site prediction inconsistencies in a cost-effective way. With a new loss, IGC encourages more penalization of the inconsistent regions during model learning and brings no extra cost to the inference.

The illustration of our IGC is presented in Fig. 3. It can be denoted as

$$v_k^* \leftarrow IGC(\hat{v}_k, D_k, \underbrace{\{\hat{v}_j | j = 1...K, j \neq k\}}_{Auxiliary Supervision}). \tag{6}$$

Concretely, given trained local models from the first stage with intermediate parameters $\hat{v}$, we first transmit all the K local models to each site, and infer them on each sample to obtain segmentation maps. Denoting the i-th sample in the k-th site as $(x_{i,k}, y_{i,k})$, we can obtain K segmentation maps produced by K local models, as $\{\hat{y}_{i,k}^j\}_{j=1}^K$. The inconsistencies are measured by computing the distance between the segmentation map inferred by the current site model and those inferred by others, as

$$\delta_{i,k} = \frac{1}{K-1} \sum_{j=1}^{K} (\hat{y}_{i,k}^k - \hat{y}_{i,k}^j)^2, \tag{7}$$

where the larger elements in the $\delta_{i,k}$ suggest the larger inconsistencies for the i-th sample. We devise a new loss function to let inconsistencies guide the model learning, increasing the supervision for the pixels with larger disagreement:

$$\begin{aligned} \mathcal{L}_{aux}(\hat{y}_{i,k}^k, y_{i,k}) = &-\text{StopGradient}(\delta_{i,k}) * \\ &[\hat{y}_{i,k}^k * \log(y_{i,k}) + (1 - \hat{y}_{i,k}^k) * \log(1 - y_{i,k})]. \end{aligned} \tag{8}$$

In addition to the weighted supervision, the general segmentation loss is also used for model optimization. The overall loss function is defined as $\mathcal{L} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{aux}$, where λ is used to balance the two objectives and is set to 1 as default.

IV. Experiments

A. Datasets and Evaluation Metrics

Datasets: We conduct the experiments on two public datasets with real-world site division and different modalities. It includes the polyp segmentation from endoscopic images (**EndoPolyp**) and the optic disc/cup segmentation from retinal fundus images (**RIF**). The detailed statistics are as follows.

- **EndoPolyp** dataset has 2187 samples collected and labeled from four different data centers [46]–[49], each of which contains $\{1000, 380, 196, 612\}$ images and labels. Here, we resize all images and labels to 384×384 and follow the train-test division used in the latest polyp segmentation work [22]. Specifically, the training sample numbers are respectively 900, 328, 170, and 550, and the rest are regarded as the test set. Since some polyp images are extracted from the same patients, we split the train-test sets in patient-level to ensure generalization.
- **RIF** dataset is built in the recent work for disc and cup segmentation in retinal images [50], whose data is collected from four different clinical sites [51]–[53]. The target region with the size of 800×800 in each image is center-cropped and resized to 384×384 following the previous work [50]. The sample number of each data site is $\{101, 159, 400, 400\}$, and in each site, the training number is $\{50, 99, 320, 320\}$, respectively. Since each image is collected from a unique patient, the image-level separation has ensured the generalization.

Metrics: We quantitatively evaluate each local site's optimized model on its test data by two commonly-used metrics, including a region-based metric, Dice score, and a boundary-based metric, average symmetric surface distance (ASSD). The larger Dice and smaller ASSD represent the better segmentation results. The detailed computation process of the prediction $\hat{y}$ and the ground-truth y is:

$$\begin{aligned} \psi_{Dice}(\hat{y}, y) &= 2 * \frac{|\hat{y} * y|}{|\hat{y}| + |y|}, \\ \psi_{ASSD}(\hat{y}, y) &= \frac{\sum_{a \in P_b} d(a, G_b) + \sum_{b \in G_b} d(b, P_b)}{|P_b| + |G_b|}, \end{aligned} \tag{9}$$

where P_b and G_b denote the predicted boundary points and the ground-truth boundary points in the $\hat{y}$ and y, and $d(\cdot, \cdot)$ denotes the minimum Euclidean distance function. Despite the two metrics being the commonly used ones to assess the segmentation results, they also have limitations in measuring the under or over-segmentation. Specifically, the Dice coefficient may not be sensitive enough to detect small under-segmentation errors or over-segmentation errors. The ASSD may be sensitive to outliers and may not accurately reflect the accuracy of the overall segmentation. Therefore, using the two metrics can provide comprehensive quantification. Moreover, as the segmentation task we investigate in this paper is binary segmentation, which does not distinguish different target objects, all target objects in an image are considered collectively as foreground pixels. The averaged scores of all local sites are used for the eventual assessment.

We understand that this limitation can impact the accuracy of our segmentation results and have discussed this in the

TABLE I

QUANTITATIVE RESULTS ON ENDOPOLYP DATASET. WE REPORT THE MEAN SCORES AND THEIR STANDARD DEVIATIONS FOR EACH SITE AS WELL AS THE AVERAGING RESULTS. THE BEST DICE (↑) AND ASSD (↓) SCORES ARE HIGHLIGHTED IN BOLD.

Site	**Dice** (%) ↑					**ASSD** ($pix.$) ↓				
	A	B	C	D	Average	A	B	C	D	Average
Local Train	75.85±23.67	57.85±20.15	29.82±23.68	83.26±22.19	61.69±20.59	17.77±17.22	22.09±17.53	47.85±55.29	10.59±18.11	24.57±14.05
FedAVG [9]	83.07±18.39	95.42±1.75	69.60±37.62	80.72±25.87	82.20±9.17	11.89±13.46	1.60±0.70	19.63±28.32	16.26±38.38	12.35±6.78
FedGKD [33]	80.02±20.30	94.92±1.91	73.53±30.48	80.55±27.39	82.26±7.82	13.71±15.89	1.76±0.73	15.91±22.14	13.68±29.33	11.27±5.56
FineTune [45]	83.07±18.40	**95.56±1.62**	69.60±37.61	82.19±24.41	82.61±9.19	11.89±13.46	**1.55±0.63**	19.61±28.30	14.65±36.90	11.93±6.60
DITTO [18]	78.61±22.10	93.47±2.24	67.41±36.89	80.82±25.20	80.08±9.25	15.02±17.15	2.91±0.78	19.38±24.95	12.53±25.71	12.46±6.03
FedRep [16]	79.62±20.38	94.03±2.00	80.69±23.79	79.63±26.99	83.49±6.10	14.86±16.77	2.38±0.62	8.72±12.24	12.91±21.08	9.72±4.78
IOP-FL [19]	79.91±21.62	95.30±1.88	77.21±26.88	81.97±23.73	83.60±6.96	13.10±16.01	1.63±0.69	12.22±16.86	14.50±36.81	10.36±5.11
FedBABU [17]	78.96±21.38	94.62±2.37	77.88±26.15	82.76±23.31	83.56±6.64	14.90±17.57	1.87±0.76	10.47±15.62	**10.33±16.47**	9.39±4.71
FedLC [5]	82.24±20.05	95.10±1.75	81.50±25.21	78.61±28.92	84.36±6.34	12.53±16.24	1.75±0.69	9.88±18.17	16.97±40.82	10.28±5.54
FedDP (Ours)	**83.21±18.52**	95.16±1.24	**84.51±19.43**	**83.45**±24.71	**86.58±4.98**	**11.03±13.15**	1.68±0.47	**7.86±10.69**	10.87±24.72	**7.86±3.78**

TABLE II

QUANTITATIVE RESULTS ON RIF DATASET. WE REPORT THE MEAN SCORES AND THEIR STANDARD DEVIATIONS FOR EACH SITE AS WELL AS THE AVERAGING RESULTS. THE BEST DICE (↑) AND ASSD (↓) SCORES ARE HIGHLIGHTED IN BOLD.

Site	**Dice** (%) ↑					**ASSD** ($pix.$) ↓				
	A	B	C	D	Average	A	B	C	D	Average
Local Train	90.31±5.87	88.69±3.72	90.24±3.38	90.53±3.00	89.94±0.73	6.25±1.93	5.70±1.53	5.39±1.48	4.53±0.82	5.47±0.62
FedAVG [9]	92.07±5.78	88.65±4.27	91.30±3.54	92.21±3.16	91.06±1.43	4.58±1.81	5.04±1.18	4.07±1.48	2.94±0.76	4.16±0.78
FedGKD [33]	91.91±5.48	88.79±3.95	91.35±3.43	92.29±3.05	91.09±1.36	4.73±1.80	4.99±1.23	4.09±1.45	2.92±0.77	4.18±0.80
FineTune [45]	92.19±6.08	89.91±4.51	91.77±3.38	92.21±3.16	91.52±0.95	4.49±1.93	4.22±1.72	3.82±1.35	2.94±0.76	3.87±0.59
DITTO [18]	92.02±5.97	90.34±4.18	91.78±3.10	92.00±3.14	91.53±0.70	4.60±1.93	4.08±1.64	3.80±1.31	3.00±0.83	3.87±0.58
FedRep [16]	92.23±5.25	89.41±4.02	91.71±3.32	92.19±3.71	91.38±1.16	4.51±1.57	4.63±1.15	3.83±1.34	2.93±0.91	3.98±0.68
IOP-FL [19]	92.42±5.36	89.34±4.16	91.68±3.65	92.52±3.57	91.49±1.28	4.22±1.74	4.49±1.20	3.61±1.52	2.48±0.80	3.70±0.77
FedBABU [17]	92.53±5.37	89.20±4.12	91.80±3.39	92.67±3.26	91.55±1.39	4.13±1.59	4.55±1.26	3.56±1.41	2.45±0.79	3.67±0.79
FedLC [5]	92.63±5.62	90.62±3.91	92.39±3.21	92.91±2.86	92.14±0.89	4.04±1.73	3.78±1.39	3.25±1.29	2.36±0.75	3.36±0.64
FedDP (Ours)	**92.96±5.80**	**91.33±3.54**	**92.46±2.98**	**93.03±2.80**	**92.44±0.68**	**3.83±1.79**	**3.42±1.56**	**3.21±1.16**	**2.33±0.73**	**3.20±0.55**

manuscript. Specifically, we have highlighted that the Dice coefficient and ASSD are not perfect metrics to quantify under and over-segmentation and have their own limitations. For example, the Dice coefficient can be biased towards larger objects, and the ASSD can be sensitive to outliers. Therefore, we have also discussed the limitations of these metrics in the manuscript and have emphasized that they should be used in conjunction with other metrics and visual inspection of the segmentation results to ensure their accuracy.

B. Implementation Details

During the dependency personalization stage, all sites adopt the same hyper-parameters. The AdamW optimizer with an initial learning rate of 0.001 is used to optimize the parameters. The gradients during training are clipped to -1 to 1 for the stable training process. Each minibatch contains six samples in all sites considering the efficiency and accuracy. For the communication round setting, enlarging the round number improves the knowledge aggregation while increasing the transmission costs. Considering the trade-off of efficiency and convergence, we empirically train the models with 200 communication rounds ($T = 200$), and the local models are trained for one epoch during each round. After that, we tune each local model for 20 epochs with a reduced learning rate of 0.0001. During training, the parameters with the best IoU score are saved for the final assessment to make the evaluation metrics more convincing and comprehensive. The whole training process is achieved on the PyTorch platform using one NVIDIA Titan 3090 GPU.

C. Comparison with State-of-the-Arts

1) Experimental Setting: We conduct the comparison to a series of GFL and PFL methods, including (a) the plain federated learning framework, FedAVG [9], (b) the latest GFL method solving data heterogeneity with no personalization, FedGKD [33], and (c) recent state-of-the-art PFL methods, i.e., FineTuning [45], DITTO [18], FedRep [16], IOP-FL [19], FedBABU [17], and FedLC [5]. We also locally train the models by only using their own datasets in different sites without FL technique as the baseline (Local Train), where we use the same self-attention-based model architecture as ours for fair comparison.

2) Quantitative Results: We report the experimental results on the EndoPolyp dataset in Table I, where our methods have achieved the best average Dice score and ASSD score, and the improvements compared to other PFL methods are obvious. Firstly, it is noticeable that all federated learning methods have gained significant improvements compared to the local training method. Particularly on Site C, as this site has less than 200 images that are not enough to train a good transformer, the results of the local training method are extremely bad and our method brings the largest performance improvement. Secondly, it is also noteworthy that the results from other FL methods on Site D are worse than the local training result. The underlying reason may be that the cases in Site D are easy to segment while the cross-site model communication harms the model learning due to the distribution difference. Despite this challenge, our method still achieves nearly the best performance verifying that our

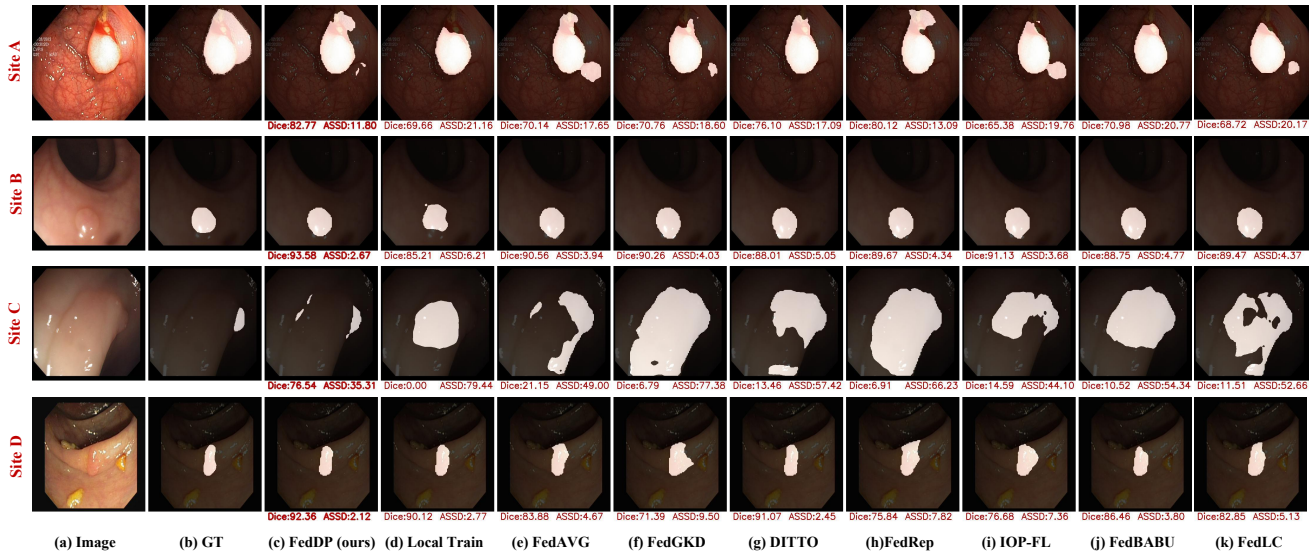


Fig. 4. Visual comparison of our method, **FedDP**, and other GFL and state-of-the-art PFL methods in EndoPolyp dataset. Each row denotes a randomly selected sample from a unique site.

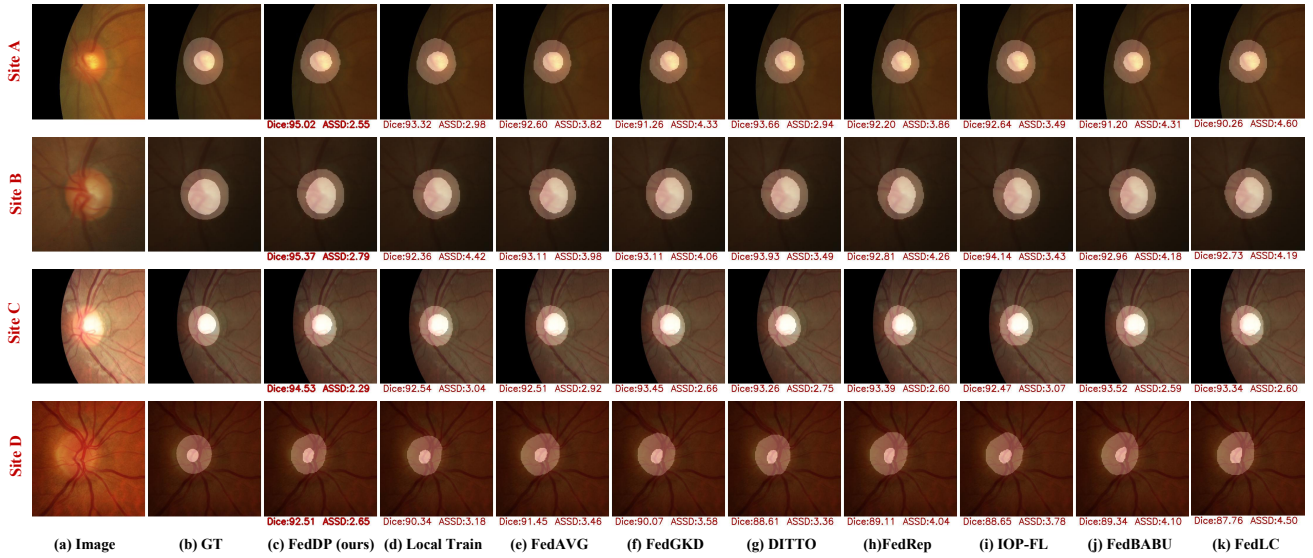


Fig. 5. Visual comparison of our method, **FedDP**, and other GFL and state-of-the-art PFL methods in RIF dataset. Each row denotes a randomly selected sample from a unique site.

dual personalization addresses the shifts well. Thirdly, it is seen that the methods exploring inter-site inconsistencies, i.e., FedLC and FedDP (ours), truly achieve better performance than others not using this information. This fact further proves that modeling the inter-site inconsistencies and taking their advantage are useful to improve local accuracy. Furthermore, our method outperforms FedLC obviously since the channel selection strategy in FedLC is designed for convolutional networks and fails to personalize the long-range dependency modeling.

The experimental results of the RIF dataset in Table II also verify the conclusions. Firstly, our method has consistently achieved the best scores on all sites. Since the heterogeneity in retinal images is not serious, all sites have gained a performance boost after the model communication, and our method results in the largest boost. Secondly, it is also observed that other PFL methods only result in limited improvements since the targets have extremely similar shapes and the slight difference locates at the boundaries. Our method improves accuracy by a large margin since our IGC can refine the predictions by encouraging the model to pay attention to analyzing site uniqueness, e.g., boundary regions.

3) Visualized Results: To further intuitively compare the performance, we visualize the segmentation results of compared methods on two datasets. For each dataset, we randomly select a sample and present its predictions as well as the ground-truth map in each site. The Dice and ASSD scores are also shown in the figure.

The results of polyp segmentation are shown in Fig. 4. For the challenging cases in Site A (the 1-st row) and Site C (the 3-rd row), nearly all other methods, including the local train, GFL, and PFL methods, have bad segmentation results owing to a large number of false detections. Instead, our method can precisely recognize the lesion position and segment the

TABLE III

ANALYTICAL ABLATION OF THE PROPOSALS, INCLUDING THE SELF-ATTENTION NETWORKS (SA), THE LOCAL QUERY (LQ), AND THE INCONSISTENCY-GUIDED CALIBRATION (IGC). THE BEST DICE (↑) AND ASSD (↓) SCORES ARE HIGHLIGHTED IN BOLD.

SA	LQ	IGC	EndoPolyp		RIF	
			Dice (%) ↑	ASSD (*pix.*) ↓	Dice (%) ↑	ASSD (*pix.*) ↓
			78.72±9.49	14.94±8.79	88.57±3.20	4.98±1.64
✓			82.20±9.17	12.35±6.78	91.06±1.43	4.16±0.78
✓	✓		84.80±5.96	8.85±4.20	92.14±0.89	3.36±0.64
✓	✓	✓	**86.58**±4.98	**7.86**±3.78	**92.44**±0.68	**3.20**±0.55

lesion boundaries with better accuracy. For the easy cases in Site B (the 2-nd row) and Site D (the 4-th row), though all methods can achieve good performance with the Dice score higher than 90%, our method improves the segmentation accuracy obviously due to its superior ability in addressing the ambiguous boundaries, which is achieved by exploring inter-site inconsistencies to enhance the perception of such regions.

The results of the disc and cup segmentation are shown in Fig. 5. Other PFL methods have not improved the segmentation accuracy obviously since they don't have the ability to perceive the boundary knowledge. Nevertheless, IGC can explore the difference in boundary predictions and take full advantage of this information to boost the results. Therefore, our method achieves the best scores on all the sites.

D. Ablation Studies

We make a detailed ablation analysis to study the effectiveness of each key component in FedDP. Then we make a series of elaborate experiments to study how each component works to improve the personalized segmentation, including the importance of changing convolutional networks into self-attention networks, personalized long-range dependency, and inconsistency-guided calibration.

1) Effectiveness of key components: We thoroughly investigate how each proposal affects each local site's segmentation accuracy by gradually advancing the baseline configuration. Specifically, the baseline method uses Feature Pyramid Network (FPN) with ResNet-18 as the architecture design and FedAVG as the federated learning algorithm. Then we gradually modify the learning program by changing ResNet-18 into the self-attention network, PVT-b0 (SA), adding the local query, and using the inconsistency-guided calibration. The results are shown in Table III. It could be seen that using self-attention networks can truly improve the scores regarding overall segmentation accuracy and boundary accuracy. Note that PVT-b0 has much fewer parameters (3.4 M) than ResNet-18 (11.7 M). Therefore, the performance improvements are attributed to the strong robustness of self-attention networks. Moreover, it is seen that LQ can further improve the metrics, i.e., 2.6% Dice score on EndoPolyp and 1.08% Dice score on RIF. IGC can further enhance the segmentation, especially in the EndoPolyp dataset, bringing a 1.78% increase in Dice and a 0.99 decrease in ASSD. The reason is that polyp lesions generally show a highly similar appearance as the surrounding tissues, and the inter-site prediction inconsistencies majorly locate in the ambiguous boundaries. IGC can distinguish the inconsistencies and make full use of this information to enhance predictions.

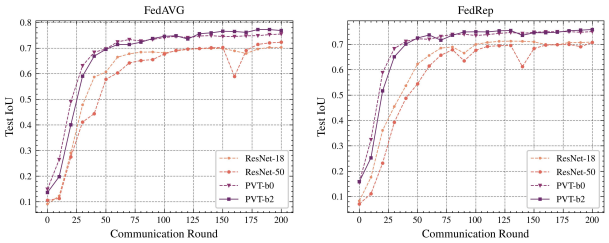


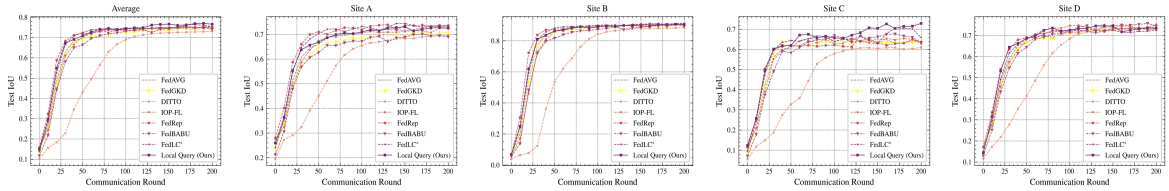
Fig. 6. Test IoU scores changing with the communication rounds on the EndoPolyp dataset using different backbones under FedAVG [9] algorithm (left) and FedRep [16] algorithm (right).

2) Self-attention Networks in Federated Segmentation: The preliminary study has proved that self-attention backbones inherently contain larger potentials in addressing the heterogeneous FL problems [27] while only considering natural image classification tasks. To additionally verify this, we replace PVT with ResNet backbones and analyze the learning process regarding convergence and performance. The experiments are conducted on EndoPolyp and we implement them on two FL algorithms. The comparison results are illustrated in Fig. 6, which show the outperforming performance of self-attention networks with faster convergence speeds and higher test scores. The results are proved consistently using two FL programs, demonstrating that the architecture of self-attention plays a key role in higher federated qualities. It is also observed that increasing the depth and channel boosts accuracy slightly. It further verifies the importance of architecture design in FL, where self-attention scheme brings main improvements compared with increasing parameters. The probable reason could be the surprisingly good robustness of capturing long-range dependency by Transformer, instead of the local contexts by convolutional networks.

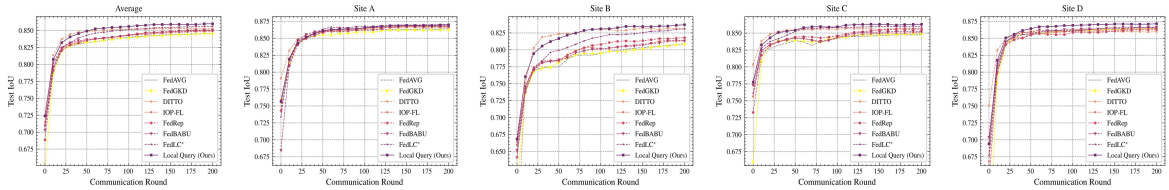
3) Long-range Dependency Personalization: Recent self-attention networks catch the long-range dependency through the **qkv** design, where **q** denotes each element's special features and **k,v** denote all elements' shared clues. Hence, our FedDP proposes to learn query embedding locally to personalize the long-range dependencies. To deeply analyze the effectiveness of this proposal, firstly we visualize the learning curves of our method with only LQ embedded and other compared methods in Fig. 7. We then show the performance of personalizing different self-attention parts in Fig. 8.

The learning curves of compared methods on two datasets are shown in Fig. 7, where FedLC* and Local Query (ours) denote the two methods exploring inter-site inconsistencies at the stage of parallel training. Firstly, the curves clearly verify that exploring the inconsistencies can improve the performance regarding the final accuracy and convergence speed. For the EndoPolyp dataset, LQ shows much faster convergence speed and higher accuracy on Site C which has a small number of samples. It supports that personalizing the query embedding process of each site truly boosts the local learning quality and enhances the segmentation accuracy of sites with limited data. For the RIF dataset, the average score curve clearly shows that LQ's performance grows obviously faster than the other methods, and of greater significance, LQ generates consistently better segmentation results on all the sites.

For analyzing the effect of personalizing different parts in

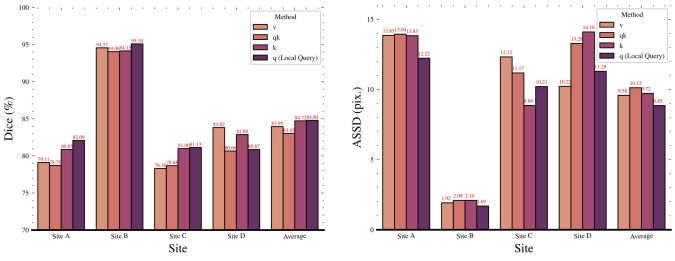


(a) Learning curves of compared methods on the EndoPolyp dataset.



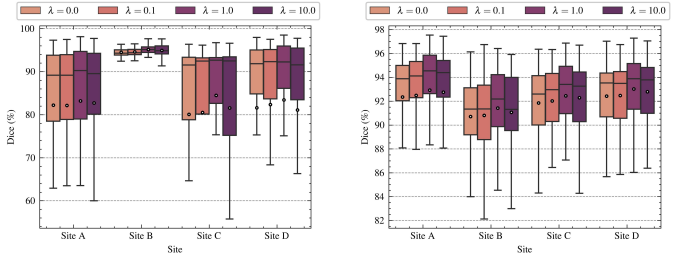
(b) Learning curves of compared methods on the RIF dataset.

Fig. 7. Test IoU scores versus the communication rounds of the average and each participated site (from left to right) on the (a) EndoPolyp dataset and the (b) RIF dataset. To simplify the learning curves and make them clearer, we present the maximum score every ten rounds.



(a) Dice scores of different personalization. **(b) ASSD scores of different personalization.**

Fig. 8. Quantitative comparison of personalizing different layers in the self-attention networks on EndoPolyp dataset. **q,k,v** denote the query, key, value embedding layers and **qk** denotes to personalize two types.



(a) Dice scores of the EndoPolyp dataset. **(b) Dice scores of the RIF dataset.**

Fig. 9. Quantitative comparison of different λ on two datasets. The white circles denote the averaged Dice scores.

qkv, the results in Fig. 8 clearly show that personalizing the query embedding layers is the most effective alternative on both metrics. Personalizing both query and key embeddings is the worst choice since it reduces too much parameter communication and thus gathers limited knowledge across sites. It is also seen that personalizing value embedding layers have poor performance since they denote the general context modeling ability better to be shared by all sites. As for the comparison of personalizing key embedding or query embedding layers, it is found that the IoU performances of the two alternatives are extremely close. The reason may be that during the calculation of **qk** similarity matrix process (Eq. 2), personalizing **q** could be equal to personalizing **k**. For the intuitive consideration that **q** denotes each element's specific features while **k** denotes the shared features, we propose to locally personalize the query embedding instead of the key embedding.

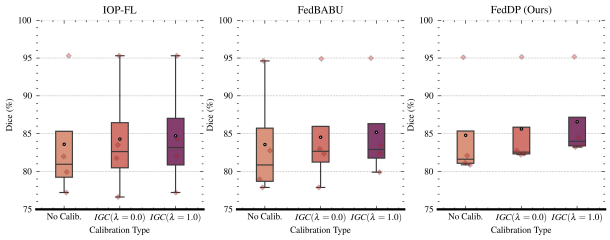


Fig. 10. Effect analysis of integrating IGC to other PFL methods, i.e., IOP-FL [19] and FedBABU [17]. The original (No Calib.) and calibrated results on EndoPolyp dataset are presented.

4) Inter-site Inconsistencies: IGC proposes to strengthen the supervision of regions with larger prediction inconsistencies. In order to balance the normal segmentation supervision ($\mathcal{L}_{seg}$) and auxiliary inconsistency-guided supervision ($\mathcal{L}_{aux}$), the combined loss objective utilizes the hype-parameter λ to control the weight. Theoretically, the larger λ lets models concentrate more on the inter-site inconsistencies and ignore fitting the ground truths to some degree. Hence, we vary this trade-off weight λ to study the effects by setting $\lambda \in \{0, 0.1, 1, 10\}$. The detailed Dice scores of two datasets are shown in Fig. 9, where we highlight the averaged scores using white circles. We can observe that the models with $\lambda \in \{0.1, 1, 10\}$ outperform the model with $\lambda = 0$, indicating that exploring inter-site inconsistencies is always beneficial to the local accuracy improvement. It is also found that model performance decreases when increasing λ from 1 to 10, verifying that too large λ negatively affects model learning with less optimization from pure segmentation loss.

Moreover, our IGC algorithm can be feasibly incorporated into other PFL methods to improve their performance. Hence, we conduct experiments to study the effects when applying IGC to refine the local models learned by other PFL algorithms, i.e., IOP-FL and FedBABU. The results on the EndoPolyp dataset are shown in Fig. 10. It is found that only using normal segmentation supervision for calibration ($\lambda = 0$) enhances the segmentation in all situations, while our IGC attains obviously larger improvements. Especially in our method, since LQ can encourage local models to learn more discriminative long-range dependencies to yield better-

TABLE IV
GENERALIZATION ANALYSIS ON UNSEEN DATA OF POLYP LESIONS. THE BEST DICE (↑) AND ASSD (↓) SCORES ARE HIGHLIGHTED IN BOLD.

Model	Dice (%) ↑				
	A	B	C	D	Ensemble
FedAVG [9]	71.24	70.93	70.81	71.84	72.10
IOP-FL [19]	71.21	71.52	71.68	72.53	72.52
FedBABU [17]	71.34	71.23	71.57	72.00	72.29
FedLC [5]	**72.27**	**72.53**	**72.85**	**72.90**	72.93
FedDP (Ours)	70.95	71.37	68.84	71.38	**73.09**

TABLE V
COMPUTATION AND TRANSMISSION EFFICIENCY ANALYSIS OF OUR METHOD AND SELECTED COMPARED METHODS.

Model	Train Time	Finetune Time	Infer. Time	Transmission (Joint, Local)
FedAVG	30 min	0 min	17.79 ms	(∼8400 MB, 0)
FedLC	30 min	3 min	18.92 ms	(∼8400 MB, 128 KB)
FedDP	30 min	3 min	17.79 ms	(∼8400 MB, 1.68 KB)

personalized features in the first learning stage, IGC can compute more valuable inconsistency knowledge to improve prediction personalization.

E. Extensive Analysis

In this part, we further study the generalization ability on unseen data sites and the computation and transmission efficiency of FedDP.

1) Generalization on Unseen Sites: As a personalized learning method, FedDP mainly aims to improve the personalized performance of the data sites participating in the federated learning framework, hence it builds a unique local model for each site and it is inevitable that there may be deviations in fitting to other unseen distributions. To quantify the generalization effects, we further collect a data site from Polyp-Gen [54] (C1, 251 images from Ambroise Paré Hospital, Paris, France) and assess the segmentation performance of the local models trained by FedAVG and different PFL methods. Table IV has shown the results where "A-D" represents the local model from Site "A" to "D" and "Ensemble" denotes the ensemble results of four local models. Since the local model has been shifted to adapt its local distribution, FedDP does not perform well when using the single local model. However, as the model ensemble method reduces the personalization effects and represents more segmentation ability, FedDP has achieved the best performance in the end, indicating its powerful and generalizable knowledge in target segmentation.

2) Efficiency Analysis: We compare the computation and transmission costs in this part to indicate the advanced efficiency of FedDP in Table V. It shows the training time, fine-tuning time, inference time, and the transmission costs at parallel training and local finetuning on the EndoPolyp dataset. Regarding computation efficiency, FedDP does not introduce any extra layers so that the inference time per image is the same as FedAVG (17.79 ms). However, the other method which explores inter-site inconsistencies costs more inference time (18.92 ms) since it models the inconsistencies through extra layers rather than the extra supervision loss in this work. As for the transmission costs, all methods nearly require the same transmission load of 8400 MB at the stage of parallel training since the personalized layers (1.6 KB in FedLC and 0.42 KB in FedDP) cover a little compared to the entire model parameters (21 MB). At the stage of local fine-tuning, since FedLC still requires the personalized layers of other sites at each training epoch, its transmission costs are much higher than FedDP, which needs the transmission only once. We also present the training and finetuning time which shows that our method spends a little more training time but gains valuable performance improvements.

V. DISCUSSION

Distribution variance widely exists in real-world multi-center resources, resulting in hard convergence and poor local performance in the GFL process. The latest researches focus on the PFL designs, allowing each participating site to save local parameters that can fit the local distribution. Nevertheless, these studies only handle the personalization of convolutional networks, ignoring the accessibility of self-attention networks. Naturally, personalized long-range dependency modeling for self-attention networks is essential, while it has not been explored in previous methods. To this end, our **FedDP** proposes LQ which decouples the query embedding layers out of the self-attention networks and saves them locally to make each local model obtain its personalized long-range modeling custom. FedDP further proposes IGC to improve prediction-level personalization. By exploring inter-site prediction-level inconsistencies, IGC integrates this inconsistency-based information into the supervision and enlarges the objectives of pixels with large inconsistencies. It facilitates local models to learn the unique prediction patterns in respective sites.

Exploring inter-site inconsistencies is exactly useful as pointed out in the previous study [5], while it suffers from heavy transmission loads. In [5], each site must obtain other sites' parameters to calculate inconsistencies in each communication round during federation. Transmission loads shall continue increasing dramatically when the site number enlarges. Instead, we compute inconsistencies using well-trained models, therefore, only require one-time communication in IGC after the first learning stage. As model parameters have been well-learned at this moment, representative inconsistency knowledge can be attained to provide precise guidance.

Considering the inter-site inconsistencies in the model learning attracts large interest in the medical domain [5], [28] owing to the special characteristics of medical data. However, most prior studies incorporate the inconsistencies into the model learning process through an extra feature-enhancing module, i.e, an inconsistency-guided attention gate [5], [28]. This is indeed helpful to refine the features but inevitably leads to more computation at the inference time. We propose to leverage inconsistencies in a computation-free manner during the inference, where the inter-site inconsistency information is integrated into the model training loss, re-weighting penalization for different pixels. Therefore, no extra computational cost is needed at the inference.

The generalization ability on unseen data sites is a valuable question to be discussed when forming a multi-site training framework. Similar to other personalized learning methods, FedDP focuses on improving the personalized performance of

each site and hence ignores the performance effects on the possible unseen data sites. On the other hand, when new sites are participating in the federated learning process or attempting to use the trained model, various methods (i.e., continuous learning, fine-tuning, and ensembling of personalized models) can be used to enhance the segmentation performance on these new data distributions. To make this paper more focused on personalized performance improvements, we will study the generalizable federated learning method in future work.

The varying number of data leads to the aggregation problem in general federated learning. This problem is mainly due to the different convergence speeds that the model trained on the fewer and easier data converges faster. To address this issue, there are typical strategies such as re-weighting the parameter aggregation by sample numbers or augmenting the data from sites with fewer samples [55]. Nevertheless, personalization encourages each local model to fit its own distribution using the local parameters, which helps avoid the negative effects of varying sample numbers of other sites.

One limitation in our method and also other PFL methods is that local personalization and global knowledge communication cannot be satisfied simultaneously. Many studies have verified that personalizing partial parameters does enhance local accuracy [14], [16], [17], while personalizing excessive parameters leads to inferior performance, such as personalizing all parameters. Therefore, it is important to control the weight between local personalization and global knowledge interaction. We plan to explore how to automatically adjust the portion of shared parameters, to maximize the information of local specialty and global universality in future work.

FedDP formulates the inconsistencies mainly regarding the spatial features and predictions since the segmentation performance relies on spatial perception. When applied to some tasks that are not sensitive to spatial knowledge, FedDP could be integrated with the channel-based methods, e.g., channel selection in FedLC, to improve the personalization ability. When processing volumetric data using 3D networks, FedDP can be easily modified to explore the 3D inconsistencies by personalizing 3D query embedding layers and calculating 3D prediction inconsistencies. As for multi-modal data, it introduces more intra-modal inconsistencies rather than inter-site inconsistencies. Leveraging the advanced modality-adaptive processing strategies in FedDP could be useful.

VI. Conclusion

This paper presents a novel personalized federated learning framework consisting of the long-range dependency personalization and prediction personalization, named as **FedDP**. It first introduces the self-attention network to the federated segmentation area and extensively improves the local performance through dual personalization. The long-range dependency personalization is achieved by the **local query** which decouples the query embedding layers out of the cross-site communication since queries denote the specialized features of each element. The **inconsistency-guided calibration** is proposed to explore the inconsistencies between different local models' predictions and use the inter-site inconsistencies to guide the local supervision. Briefly, the supervision is modified by adding an auxiliary objective that enlarges the attention of pixels with large inconsistencies. The effectiveness of our method is well verified on two medical image segmentation tasks with detailed ablation analysis. Besides the accuracy improvement, our method has large potential in practical application thanks to its cost-effective design.

VII. Acknowledgement

This work is supported by the Ministry of Science and Technology of the People's Republic of China under grant No. 2021ZD0201900 and 2021ZD0201904, and supported by WEISS [203145/Z/16/Z]; and Horizon 2020 FET (863146). For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

References

[1] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[2] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020.

[3] W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M. J. Cardoso *et al.*, "Privacy-preserving federated brain tumour segmentation," in *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer, 2019, pp. 133–141.

[4] S. Pati, U. Baid, M. Zenk, B. Edwards, M. Sheller, G. A. Reina, P. Foley, A. Gruzdev, J. Martin, S. Albarqouni *et al.*, "The federated tumor segmentation (fets) challenge," *arXiv preprint arXiv:2105.05874*, 2021.

[5] J. Wang, Y. Jin, and L. Wang, "Personalizing federated medical image segmentation via local calibration," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*. Springer, 2022, pp. 456–472.

[6] H. R. Roth, D. Yang, W. Li, A. Myronenko, W. Zhu, Z. Xu, X. Wang, and D. Xu, "Federated whole prostate segmentation in mri with personalized neural architectures," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 357–366.

[7] X. Xu and P. Yan, "Federated multi-organ segmentation with partially labeled data," *arXiv preprint arXiv:2206.07156*, 2022.

[8] P. Liu, M. Sun, and S. K. Zhou, "Multi-site organ segmentation with federated partial supervision and site adaptation," *arXiv preprint arXiv:2302.03911*, 2023.

[9] 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.

[10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

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

[12] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5650–5659.

[13] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *arXiv preprint arXiv:1911.07963*, 2019.

[14] 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.

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

[16] 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.

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

[18] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6357–6368.

[19] M. Jiang, H. Yang, C. Cheng, and Q. Dou, "Iop-fl: Inside-outside personalization for federated medical image segmentation," *arXiv preprint arXiv:2204.08467*, 2022.

[20] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201*, 2021.

[21] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-aware transformers for skin lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 206–216.

[22] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," *arXiv preprint arXiv:2108.06932*, 2021.

[23] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873*, 2022.

[24] J. Wang, F. Chen, Y. Ma, L. Wang, Z. Fei, J. Shuai, X. Tang, Q. Zhou, and J. Qin, "Xbound-former: Toward cross-scale boundary modeling in transformers," *IEEE Transactions on Medical Imaging*, 2023.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1439–1449.

[27] L. Qu, Y. Zhou, P. P. Liang, Y. Xia, F. Wang, E. Adeli, L. Fei-Fei, and D. Rubin, "Rethinking architecture design for tackling data heterogeneity in federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 061–10 071.

[28] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.

[29] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[30] T. Yu, E. Bagdasaryan, and V. Shmatikov, "Salvaging federated learning by local adaptation," *arXiv preprint arXiv:2002.04758*, 2020.

[31] L. Zhang, L. Shen, L. Ding, D. Tao, and L.-Y. Duan, "Fine-tuning global model via data-free knowledge distillation for non-iid federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 174–10 183.

[32] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3557–3568, 2020.

[33] D. Yao, W. Pan, Y. Dai, Y. Wan, X. Ding, H. Jin, Z. Xu, and L. Sun, "Local-global knowledge distillation in heterogeneous federated learning with non-iid data," *arXiv preprint arXiv:2107.00051*, 2021.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[37] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[38] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[39] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[41] Z. Dai, B. Cai, Y. Lin, and J. Chen, "Up-detr: Unsupervised pre-training for object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1601–1610.

[42] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 171–180.

[43] 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.

[44] 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.

[45] K. Wang, R. Mathews, C. Kiddon, H. Eichner, F. Beaufays, and D. Ramage, "Federated evaluation of on-device personalization," *arXiv preprint arXiv:1910.10252*, 2019.

[46] 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.

[47] 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, no. 2, pp. 283–293, 2014.

[48] 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.

[49] 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 *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[50] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1013–1023.

[51] F. J. F. Batista, T. Diaz-Aleman, J. Sigut, S. Alayon, R. Arnay, and D. Angel-Pereira, "Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning," *Image Analysis & Stereology*, vol. 39, no. 3, pp. 161–167, 2020. [Online]. Available: https://www.ias-iss.org/ojs/IAS/article/view/2346

[52] 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.

[53] 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.

[54] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, M. A. Riegler, K. V. Anonsen *et al.*, "A multi-centre polyp detection and segmentation dataset for generalisability assessment," *Scientific Data*, vol. 10, no. 1, p. 75, 2023.

[55] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang, "Self-balancing federated learning with global imbalanced data in mobile systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 59–71, 2020.