

FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space

Quande Liu¹, Cheng Chen¹, Jing Qin², Qi Dou^{1,*}, Pheng-Ann Heng¹

¹ Department of Computer Science and Engineering, The Chinese University of Hong Kong

² School of Nursing, The Hong Kong Polytechnic University

{qdliu, cchen, qdou, pheng}@cse.cuhk.edu.hk, harry.qin@polyu.edu.hk

Abstract

Federated learning allows distributed medical institutions to collaboratively learn a shared prediction model with privacy protection. While at clinical deployment, the models trained in federated learning can still suffer from performance drop when applied to completely unseen hospitals outside the federation. In this paper, we point out and solve a novel problem setting of federated domain generalization (FedDG), which aims to learn a federated model from multiple distributed source domains such that it can directly generalize to unseen target domains. We present a novel approach, named as Episodic Learning in Continuous Frequency Space (ELCFS), for this problem by enabling each client to exploit multi-source data distributions under the challenging constraint of data decentralization. Our approach transmits the distribution information across clients in a privacy-protecting way through an effective continuous frequency space interpolation mechanism. With the transferred multi-source distributions, we further carefully design a boundary-oriented episodic learning paradigm to expose the local learning to domain distribution shifts and particularly meet the challenges of model generalization in medical image segmentation scenario. The effectiveness of our method is demonstrated with superior performance over state-of-the-arts and in-depth ablation experiments on two medical image segmentation tasks. The code is available at <https://github.com/liuquande/FedDG-ELCFS>.

1. Introduction

Data collaboration across multiple medical institutions is increasingly desired to build accurate and robust data-driven deep networks for medical image segmentation [7, 18, 50]. Federated learning (FL) [20] has recently opened the door for a promising privacy-preserving solution, which allows training a model on distributed datasets while keeping data

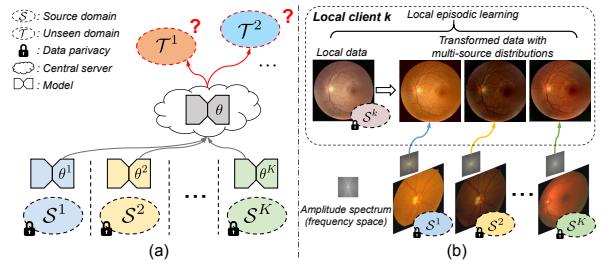


Figure 1. (a) The novel problem setting of federated domain generalization (FedDG), which aims to learn a federated model from multiple decentralized source domains such that it can directly generalize to completely unseen target domains. (b) Our main idea to tackle FedDG by transferring distribution information in frequency space and episodic learning at each local client.

locally. The paradigm works in a way that each local client (e.g., hospital) learns from their own data, and only aggregates the model parameters at a certain frequency at the central server to generate a global model. All the data samples are kept within each local client during federated training.

Although FL has witnessed some pilot progress on medical image segmentation tasks [4, 44, 49], all existing works only focus on improving model performance on the internal clients, while neglecting model generalizability onto unseen domains outside the federation. This is a crucial problem impeding wide applicability of FL models in real practice. The testing medical images encountered in unseen hospitals can differ significantly from the source clients in terms of data distributions, due to the variations in imaging scanners and protocols. How to generalize the federated model under such distribution shifts is technically challenging yet unexplored so far. In this work, we identify the novel problem setting of *Federated Domain Generalization* (FedDG), which aims to learn a federated model from multiple decentralized source domains such that it can directly generalize to completely unseen domains, as illustrated in Fig. 1 (a).

Unseen domain generalization (DG) is an active research topic with different methods being proposed [3, 8, 11, 24, 25, 26, 29, 37, 43], but the federated paradigm with dis-

*Corresponding author

tributed data sources poses new challenges for DG. With the goal to extract representations that are robust to distribution shift, existing DG approaches usually require access to multi-source distributions in the learning process. For instance, adversarial feature alignment methods [26, 29] have to train the domain discriminator with samples from different source datasets. Meta-learning based methods [8, 24] need to use multi-source data of different distributions to construct virtual training and virtual testing domains within each minibatch. Whereas in federated paradigm, data are stored distributedly and *the learning at each client can only access its local data*. Therefore, current DG methods are typically not applicable in FedDG scenario. In addition, the local optimization would make model biased to its own data distribution, thus less generalizable to new target domains.

To solve this FedDG problem, our insight is to enable each client to access multi-source data distributions in a privacy-protecting way. The idea is motivated by the knowledge that the low-level distributions (i.e., style) and high-level semantics of an image can be respectively captured by amplitude and phase spectrum in the frequency space, as revealed by visual psychophysics [13, 42, 57]. We can consider exchanging these amplitude spectrum across clients to transmit the distribution information (cf. Fig. 1 (b)), while keeping the phase spectrum with core semantics locally for privacy protection. Based on this, we also devise a continuous frequency space interpolation mechanism, which interpolates between the local and transferred distributions for enriching the established multi-domain distributions for each local client. This promotes the local training to gain domain-invariance benefiting from a dedicated dense distribution space. With these established distributions, we expose the local learning to domain distribution shifts via an episodic training paradigm to enhance the generalizability of local parameters. A novel meta-update objective function is designed to guide cross-domain optimization attending to the boundary area. This is notably important for medical image segmentation applications where generalization errors often come from imprecise predictions at ambiguous boundary of anatomies.

Our main contributions are highlighted as follows:

- We tackle the novel and practical problem of *Federated Domain Generalization*. To the best of our knowledge, this is the first work to improve generalizability on completely unseen domains for federated models.
- We propose a privacy-preserving solution to learn the generalizable FL model under decentralized datasets, through an effective continuous frequency space interpolation mechanism across clients.
- We present a novel boundary-oriented episodic learning scheme for the local training at a client, which exposes local optimization to domain shifts and enhances

model generalizability at ambiguous boundary area.

- We conduct extensive experiments on two typical medical image segmentation tasks, i.e., retinal fundus image segmentation (four datasets) and prostate MRI segmentation (six datasets). Our achieved superior performance over state-of-the-arts and in-depth analytical experiments demonstrate the efficacy of our approach.

2. Related Work

2.1. Federated Learning in Medical Imaging

Federated learning [15, 20, 36, 56] provides a promising privacy-preserving solution for multi-site data collaboration, which develops a global model from decentralized datasets by aggregating the parameters of each local client while keeping data locally. Representatively, McMahan et al. [36] propose the popular federated averaging algorithm for communication-efficient federated training of deep networks. With the advantage of privacy protection, FL has recently drawn increasing interests in medical image applications [4, 18, 22, 27, 45, 49, 51]. Sheller et al. [49] is a pilot study to investigate the collaborative model training without sharing patient data for the multi-site brain tumor segmentation. Later on, Li et al. [27] further compare several weights sharing strategies in FL to alleviate the effect of data imbalance among different hospitals. However, these works all focus on improving performance on internal clients, without considering the generalization issue for unseen domains outside the federation, which is crucial for wide clinical usability. Latest literature has studied a related problem of unsupervised domain adaptation in FL paradigm [28, 41], whereas these methods typically require data from the target domain to adapt the model. In practice, it would be time-consuming or even impractical to collect data from each new hospital before model deployment. Instead, our tackled new problem setting of FedDG aims to directly generalize the federated model to completely unseen domains, in which no prior knowledge from the target domain is needed.

2.2. Domain Generalization

Domain generalization [5, 9, 12, 14, 43, 47, 58, 59] aims to learn a model from multiple source domains such that it can directly generalize to unseen target domains. Among previous efforts, some methods aim to learn domain-invariant representations by minimizing the domain discrepancy across multiple source domains [11, 16, 26, 29, 32, 37, 38, 55]. For example, Motiian et al. [37] utilize a contrastive loss to minimize the distance between samples from the same class but different domains. Some other DG methods are based on meta-learning, which is an episodic training paradigm by creating meta-train and meta-test splits at each iteration to stimulate domain shift [1, 8, 24, 30]. Li et al. [30] employ meta-learning

to learn an auxiliary loss that guides the feature extractor to learn more generalized features. However, these methods typically require centralizing multi-domain data in one place for learning, which violates privacy protection in federated learning setting with decentralized datasets.

There are other methods tackling DG by manipulating deep neural network architectures [19, 23, 35], leveraging self-supervision signals [3, 54], designing training heuristics [17, 25], or conducting data augmentations [48, 53, 60, 61], which are free from requirement of data centralization. Representative, Carlucci et al. [3] adopt self-supervised learning by solving jigsaw puzzles. Zhang et al. [60] conduct extensive data augmentations on each source domain by stacking a series of transformations. These approaches, when applied in FL paradigm, can helpfully act as regularizations for the local training with individual source domain data, yet hardly exploit the rich data distributions across domains. Our method instead, aims to transfer the distribution information across clients to make full use of the multi-source distributions towards FedDG. We also experimentally compare with these typical methods under the FL setting with superior performance demonstrated.

3. Method

We start with the formulation for federated domain generalization and its challenges in medical image segmentation scenario. We then describe the proposed method *Episodic Learning in Continuous Frequency Space* (EL-CFS) to explicitly address these challenges. An overview of the method is shown in Fig. 2.

3.1. Federated Domain Generalization

Preliminaries: In FedDG, we denote $(\mathcal{X}, \mathcal{Y})$ as the joint image and label space of a task, $\mathcal{S} = \{\mathcal{S}^1, \mathcal{S}^2, \dots, \mathcal{S}^K\}$ as the set of K distributed source domains involved in federated learning. Each domain contains data and label pairs of $\mathcal{S}^k = \{(x_i^k, y_i^k)\}_{i=1}^{N^k}$, which are sampled from a domain-specific distribution $(\mathcal{X}^k, \mathcal{Y})$. The goal of FedDG is to learn a model $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ using the K distributed source domains, such that it can directly generalize to a completely unseen testing domain \mathcal{T} with a high performance.

Standard federated learning paradigm involves the communication between a central server and the K local clients. At each federated round t , every client k will receive the same global model weights θ from the central server and update the model with their local data \mathcal{S}^k for E epochs. The central server then collects the local parameters θ^k from all clients and aggregates them to update the global model. This process repeats until the global model converges. In this work, we consider the most popular federated averaging algorithm (FedAvg) [36], which aggregates the local parameters with weights in proportional to the size of each local

dataset to update the global model, i.e., $\theta = \sum_{k=1}^K \frac{N^k}{N} \theta^k$, where $N = \sum_{k=1}^K N^k$. It is worth noting that our method can also be flexibly incorporated to other FL backbones.

Challenges: With the goal of unseen domain generalization, a model is expected to thoroughly investigate the multi-source data distributions to pursue domain-invariance of its learned latent space. However, the federated setting in the specific medical image segmentation scenario poses several challenges for that. *First*, the multi-source data in FL are stored distributedly and the learning at each client can only access its individual local distribution, which constrains to make full use of the multi-source distributions to learn generalizable parameters. *Second*, though FL has collaborated multi-source data, the medical images acquired from different clinical sites can present large heterogeneity. This leads to distinct distributions among the collaborative datasets, which is insufficient to ensure domain invariance in a more continuous distribution space to attain good generalizability in complex clinical environments. *Third*, the structure of medical anatomises usually present high ambiguity around its boundary region, raising challenge for previous DG techniques that typically lacks assurance for the domain-invariance of features in such ambiguous region.

3.2. Continuous Frequency Space Interpolation

To address the restriction of decentralized datasets, the foundation of our solution is to exchange the distribution information across clients, such that each local client can get access to multi-source data distributions for learning generalizable parameters. Considering that sharing raw images is forbidden, we propose to exploit the information inherent in the frequency space, which enables to separate the distribution (i.e. style) information from the original images to be shared between clients without privacy leakage.

Specifically, given a sample $x_i^k \in \mathbb{R}^{H \times W \times C}$ ($C = 3$ for RGB image and $C = 1$ for grey-scale image) from the k -th client, we can obtain its frequency space signal through fast Fourier transform [39] as:

$$\mathcal{F}(x_i^k)(u, v, c) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x_i^k(h, w, c) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)}. \quad (1)$$

This frequency space signal $\mathcal{F}(x_i^k)$ can be further decomposed to an amplitude spectrum $\mathcal{A}_i^k \in \mathbb{R}^{H \times W \times C}$ and a phase spectrum $\mathcal{P}_i^k \in \mathbb{R}^{H \times W \times C}$, which respectively reflect the low-level distributions (e.g. style) and high-level semantics (e.g. object) of the image. To exchange the distribution information across clients, we first construct a distribution bank $\mathcal{A} = [\mathcal{A}^1, \dots, \mathcal{A}^K]$, where each $\mathcal{A}^k = \{\mathcal{A}_i^k\}_{i=1}^{N^k}$ contains all amplitude spectrum of images from the k -th client, representing the distribution of \mathcal{X}^k . This bank is then made accessible to all clients as shared distribution knowledge.

Next, we design a continuous interpolation mechanism within the frequency space, aiming to transmit multi-source

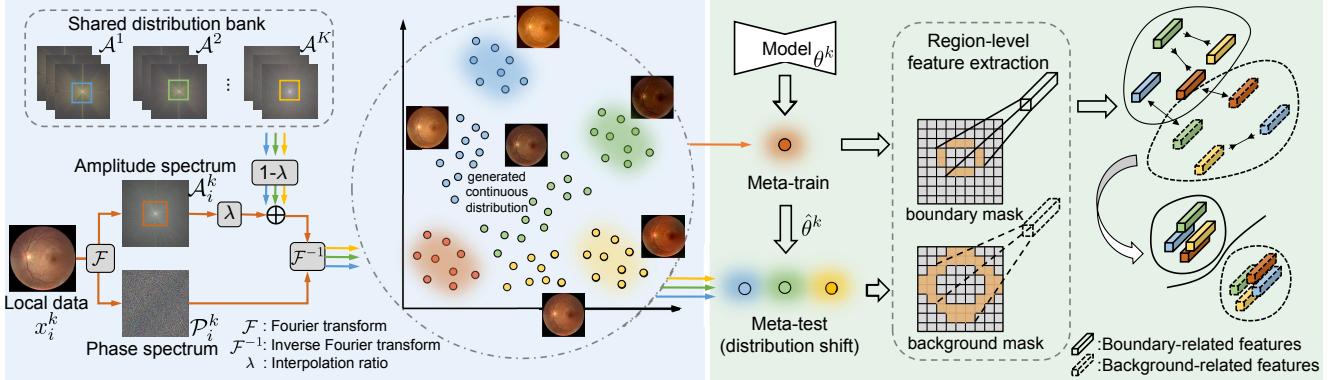


Figure 2. Overview of our proposed episodic learning in continuous frequency space (ELCFS). The distribution information is exchanged across clients from frequency space with a continuous interpolation mechanism, enabling each local client to access the multi-source distributions. An episodic training paradigm is then established to expose the local optimization to domain shift, with explicit regularization to promote domain-independent feature cohesion and separation at the ambiguous boundary region for improving generalizability.

distribution information to a local client leveraging the distribution bank. As shown in the left part of Fig. 2, given a local image x_i^k at client k , we can replace some low-frequency component of its amplitude spectrum with the ones in distribution bank \mathcal{A} , while its phase spectrum is unaffected to preserve the semantic content. As an outcome, we can generate images with transformed appearances exhibiting distribution characteristics of other clients. More importantly, we continuously interpolate between amplitude spectrum of local data and the transferred amplitude spectrum of other domains. In this way, we can enrich the established multi-domain distributions for each local client, benefiting from a dedicated dense space with smooth distribution changes. Formally, this is achieved by randomly sampling an amplitude spectrum item $\mathcal{A}_j^n (n \neq k)$ from the distribution bank, then synthesize a new amplitude spectrum by interpolating between \mathcal{A}_i^k and \mathcal{A}_j^n . Let $\mathcal{M} = \mathbb{1}_{(h,w) \in [-\alpha H : \alpha H, -\alpha W : \alpha W]}$ be a binary mask which controls the scale of low-frequency component within amplitude spectrum to be exchanged, whose value is 1 at the central region and 0 elsewhere. Denote λ as the interpolation ratio adjusting the amount of distribution information contributed by \mathcal{A}_i^k and \mathcal{A}_j^n , the generated new amplitude spectrum interacting distributions for local client k and external client n is represented as:

$$\mathcal{A}_{i,\lambda}^{k \rightarrow n} = (1 - \lambda)\mathcal{A}_i^k * (1 - \mathcal{M}) + \lambda\mathcal{A}_j^n * \mathcal{M}. \quad (2)$$

After obtaining the interpolated amplitude spectrum $\mathcal{A}_{i,\lambda}^{k \rightarrow n}$, we then combine it with the original phase spectrum to generate the transformed image via inverse Fourier transform \mathcal{F}^{-1} as follows:

$$x_{i,\lambda}^{k \rightarrow n} = \mathcal{F}^{-1}(\mathcal{A}_{i,\lambda}^{k \rightarrow n}, \mathcal{P}_i^k), \quad (3)$$

where the generated image $x_{i,\lambda}^{k \rightarrow n}$ preserves the original semantics of x_i^k while carrying a new distribution interacted between \mathcal{X}^k and \mathcal{X}^n . In our implementation, the interpolation ratio λ will be dynamically sampled from [0, 1] to generate images via a continuous distribution space.

As intuitive examples shown in Fig. 2, our interpolation operation allows the generated samples to bridge the intermediate space between distinct distributions across domains. Note that the method described above does not require heavy computations, thus can be performed online as the local learning goes on. Practically, for each input x_i^k , we will sample an amplitude spectrum \mathcal{A}_j^n from the distribution bank for each external client $n \neq k$, and transform its image appearance as Eqs. (2-3). Through this, we obtain $K-1$ transformed images $\{x_{i,\lambda}^{k \rightarrow n}\}_{n \neq k}$ of different distributions, which share the same semantic label as x_i^k . For ease of denotation, we represent these transformed images as t_i^k hereafter, i.e. $t_i^k = \{x_{i,\lambda}^{k \rightarrow n}\}_{n \neq k}$. Furthermore, this approach does not violate the privacy concern since the phase spectrum containing core semantics are retained at each client throughout the whole process, and the raw images cannot be reconstructed with the amplitude spectrum alone [46].

3.3. Boundary-oriented Episodic Learning

The above constructed continuous multi-source distributions at each local client provide the materials to learn generalizable local parameters. In the following, we carefully design a boundary-oriented episodic learning scheme for local training, by particularly meeting challenges of model generalization in medical image segmentation scenario.

Episodic learning at local client: We establish the local training as an episodic meta-learning scheme, which learns generalizable model parameters by simulating train/test domain shift explicitly. Note that in our case, the domain shift at a local client comes from the data generated from frequency space with different distributions. Specifically, in each iteration, we consider the raw input x_i^k as meta-train and its counterparts t_i^k generated from frequency space as meta-test presenting distribution shift (cf. Fig. 2). The

meta-learning scheme can then be decoupled to two steps. First, the model parameters θ^k are updated on meta-train with segmentation Dice loss \mathcal{L}_{seg} :

$$\hat{\theta}^k = \theta^k - \beta \nabla_{\theta^k} \mathcal{L}_{seg}(x_i^k; \theta^k), \quad (4)$$

where β denotes the learning rate for the inner-loop update. Second, a meta-update is performed to virtually evaluate the updated parameters $\hat{\theta}^k$ on the held-out meta-test data t_i^k with a meta-objective \mathcal{L}_{meta} . Crucially, this objective is computed with the updated parameters $\hat{\theta}^k$, but optimized w.r.t the original parameters θ^k . Such optimization paradigm aims to train the model such that its learning on source domains can further fulfill certain properties that we desire in unseen domains, which are quantified by \mathcal{L}_{meta} .

Boundary-oriented meta optimization: We define the \mathcal{L}_{meta} with considering specific challenges in medical image segmentation. Particularly, it is observed that the performance drop of segmentation results at unseen domains outside federation often comes from the ambiguous boundary area of anatomies. To this end, we design a new boundary-oriented objective to enhance the domain-invariant boundary delineation, by carefully learning from the local data x_i^k and the corresponding t_i^k generated from frequency space with multi-source distributions. The idea is to regularize the boundary-related and background-related features of these data to respectively cluster to a compact space regardless of their distributions while reducing the clusters overlap. This is crucial, since if the model cannot project their features around boundary area with distribution-independent class-specific cohesion and separation, the predictions will suffer from ambiguous decision boundaries and still be sensitive to the distribution shift when deployed to unseen domains outside federation.

Specifically, we first extract the boundary-related and background-related features for the input samples. Given image x_i^k with segmentation label y_i^k , we can extract its binary boundary mask $y_{i,bd}^k$ and background mask $y_{i,bg}^k$ with morphological operations on y_i^k . Here, the mask $y_{i,bg}^k$ only contains background pixels around the anatomy boundary instead of from the whole image, as we expect to enhance the discriminability for features around the boundary region. Let Z_i^k denote the activation map extracted from layer l , which is interpolated with bilinear interpolation to keep consistent dimensions as y_i^k . Then the boundary-related and background-related features of x_i^k can be extracted from Z_i^k with masked average pooling over $y_{i,bd}^k$ and $y_{i,bg}^k$ as:

$$h_{i,bd}^k = \frac{\sum_{h,w} Z_i^k * y_{i,bd}^k}{\sum_{h,w} y_{i,bd}^k}; h_{i,bg}^k = \frac{\sum_{h,w} Z_i^k * y_{i,bg}^k}{\sum_{h,w} y_{i,bg}^k}, \quad (5)$$

where $*$ denote element-wise product. The produced $h_{i,bd}^k$ and $h_{i,bg}^k$ are single-dimensional vectors, representing the averaged region-level features of the boundary and background pixels. By further performing the same operation

for $K-1$ transformed images t_i^k with different distributions transferred from the frequency space, we accordingly obtain together K boundary-related and K background-related features.

Next, we enhance the domain-invariance and discriminability of these features, by regularizing their intra-class cohesion and inter-class separation regardless of distributions. Here, we employ the well-established InfoNCE [6] objective to impose such regularization. Denote (h_m, h_p) as a pair of features, which is a positive pair if h_m and h_p are of the same class (both boundary-related or background-related) and otherwise negative pair. In our case, the InfoNCE loss is defined over each positive pair (h_m, h_p) within the $2 \times K$ region-level features as:

$$\ell(h_m, h_p) = -\log \frac{\exp(h_m \odot h_p / \tau)}{\sum_{q=1, q \neq m}^{2K} \mathbb{F}(h_m, h_q) \cdot \exp(h_m \odot h_q / \tau)}, \quad (6)$$

where \odot denote the cosine similarity: $a \odot b = \frac{\langle a, b \rangle}{\|a\|_2 \|b\|_2}$; the value of $\mathbb{F}(h_m, h_q)$ is 0 and 1 for positive and negative pair respectively; τ denotes the temperature parameter. The final loss $\mathcal{L}_{boundary}$ is the average of ℓ over all positive pairs:

$$\mathcal{L}_{boundary} = \sum_{m=1}^{2K} \sum_{p=m+1}^{2K} \frac{(1 - \mathbb{F}(h_m, h_p)) \cdot \ell(h_m, h_p)}{B(K, 2) \times 2}, \quad (7)$$

where $B(K, 2)$ is the number of combinations.

Overall local learning objective: The overall meta objective is composed of the segmentation dice loss \mathcal{L}_{seg} and the boundary-oriented objective $\mathcal{L}_{boundary}$ as:

$$\mathcal{L}_{meta} = \mathcal{L}_{seg}(t_i^k; \hat{\theta}^k) + \gamma \mathcal{L}_{boundary}(x_i^k, t_i^k; \hat{\theta}^k), \quad (8)$$

where $\hat{\theta}^k$ is the updated parameter from Eq. 4, γ is a balancing hyper-parameter. Finally, both the inner-loop objective and meta objective will be optimized together with respect to the original parameter θ^k as:

$$\arg \min_{\theta^k} \mathcal{L}_{seg}(x_i^k; \theta^k) + \mathcal{L}_{meta}(x_i^k, t_i^k; \hat{\theta}^k). \quad (9)$$

In a federated round, once the local learning is finished, the local parameters θ^k from all clients will be aggregated at the central server to update the global model.

4. Experiments

We extensively evaluate our method on two medical image segmentation tasks, i.e., the optic disc and cup segmentation on retinal fundus images [40], and the prostate segmentation on T2-weighted MRI [31]. We first conduct comparison with DG methods that can be incorporated in the federated paradigm, and then provide in-depth ablation studies to analyze our method.

4.1. Datasets and Evaluation Metrics

We employ **retinal fundus images from 4 different clinical centers** of public datasets [52, 10, 40] for optic disc and cup segmentation. For pre-processing, we center-crop a 800×800 disc region for these data uniformly, then resize the cropped region to 384×384 as network input. We further collect **prostate T2-weighted MRI images from 6 different data sources** partitioned from the public datasets [2, 21, 31, 33] for prostate MRI segmentation task. All the data are pre-processed to have similar field of view for the prostate region and resized to 384×384 in axial plane. We then normalize the data individually to zero mean and unit variance in intensity values. Note that for both tasks, the data acquired from different clinical centers present heterogeneous distributions due to the varying imaging conditions. The example cases and sample numbers of each data source are presented in Fig. 3. Data augmentation of random rotation, scaling, and flipping are employed in the two tasks. For evaluation, we adopt two commonly-used metrics of Dice coefficient (Dice) and Hausdorff distance (HD), to quantitatively evaluate the segmentation results on the whole object region and the surface shape respectively.

4.2. Implementation Details

In the federated learning process, all clients use the same hyper-parameter settings, and the local model is trained using Adam optimizer with batch size of 5 and Adam momentum of 0.9 and 0.99. The meta-step size and learning rate are both set as $1e^{-3}$. The interpolation ratio λ in frequency space is randomly sampled within [0.0, 1.0], and we will investigate this parameter in the ablation study. The hyper-parameter α is empirically set as 0.01 to avoid artifacts on the transformed images. The activation map from the last two deconvolutional layers are interpolated and concatenated to extract the semantic features around boundary region, and the temperature parameter τ is empirically set as 0.05. The weight γ is set as 0.1 and 0.5 in the two tasks to balance the magnitude of the training objectives. We totally train 100 federated rounds as the global model has converged stably, and the local epoch E in each federated round is set as 1. The framework is implemented with Pytorch library, and is trained on two NVIDIA TitanXp GPUs.

4.3. Comparison with DG methods

Experimental setting: In our experiments, we follow the practice in domain generalization literature to adopt the leave-one-domain-out strategy, i.e., training on $K-1$ distributed source domains and testing on the one left-out unseen target domain. This results in four generalization settings for the fundus image segmentation task and six settings for the prostate MRI segmentation task.

We compare with recent state-of-the-art DG methods that are free from data centralization and can be incorpo-

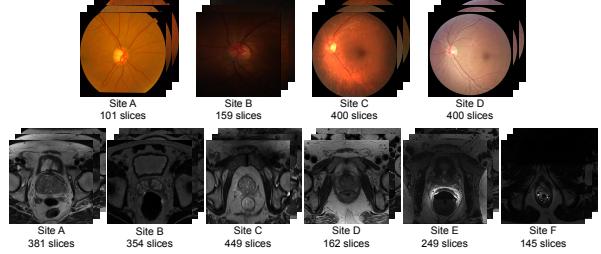


Figure 3. Example cases and slice number of each data source in fundus image segmentation and prostate MRI segmentation tasks.

rated into the local learning process in federated paradigm, including: **JiGen** [3] an effective self-supervised learning approach to learn general representations by solving jigsaw puzzles; **BigAug** [60] a method that performs extensive data transformations to regularize general representation learning; **Epi-FCR** [25] a scheme to periodically exchange partial model (classifier or feature extractor) across domains to expose model learning to domain shift; **RSC** [17] a method that randomly discards the dominating features to promote robust model optimization. For the implementation, we follow their public code or paper and establish them in the federated setting. We also compare with the baseline setting, i.e., learning a global model with the basic **FedAvg** [36] algorithm without any generalization technique.

Comparison results: Table 1 presents the quantitative results for retinal fundus segmentation. We see that different DG methods can improve the overall generalization performance more or less over FedAvg. This attributes to their regularization effect on the local learning to extract general representations. Compared with these methods, our ELCFS achieves higher overall performance and obtains improvements on most unseen sites in terms of Dice and HD for both optic disc and cup segmentation. This benefits from our frequency space interpolation mechanism which presents multi-domain distributions to local client. Specifically, for other DG methods, their local learning still can only access the individual distribution and fail to regularize the features towards domain-invariance in a diverse distribution space. In contrast, our method enables the local learning to take full advantages of the multi-source distributions and explicitly enhances the domain-invariance of features around the ambiguous boundary region. In addition, our ELCFS achieves consistent improvements over FedAvg across all unseen domain settings, with the overall performance increase of 2.02% in Dice and 2.86 in HD. For prostate MRI segmentation, the comparison DG methods generally perform better than FedAvg, but the improvements are relatively marginal. Our ELCFS obtains the highest Dice across all the six unseen sites and HD on most sites. Overall, our method improves over FedAvg for Dice from 85.57% to 87.39% and HD from 12.42 to 10.88, outperforming other DG methods. Fig. 4 shows the segmentation

Table 1. Comparison of federated domain generalization results on Optic Disc/Cup segmentation from fundus images.

Task	Optic Disc Segmentation					Optic Cup Segmentation					Overall	Optic Disc Segmentation					Optic Cup Segmentation					Overall	
	A	B	C	D	Avg.	A	B	C	D	Avg.		A	B	C	D	Avg.	A	B	C	D	Avg.		
Dice Coefficient (Dice) \uparrow												Hausdorff Distance (HD) \downarrow											
JiGen [3]	93.92	85.91	92.63	94.03	91.62	82.26	70.68	83.32	85.70	80.47	86.06	13.12	20.18	11.29	8.15	13.19	20.88	23.21	11.55	9.23	16.22	14.71	
BigAug [60]	93.49	86.18	92.09	93.67	91.36	81.62	69.46	82.64	84.51	79.56	85.46	16.91	19.01	11.53	8.76	14.05	21.21	23.10	12.02	10.47	16.70	15.39	
Epi-FCR [25]	94.34	86.22	92.88	93.73	91.79	83.06	70.25	83.68	83.14	80.03	85.91	13.02	18.97	10.67	8.47	12.78	19.12	21.94	11.50	10.86	15.86	14.32	
RSC [17]	94.50	86.21	92.23	94.15	91.77	81.77	69.37	83.40	84.82	79.84	85.80	19.44	19.26	13.47	8.14	15.08	23.85	24.01	11.38	9.79	17.25	16.16	
FedAvg [36]	92.88	85.73	92.07	93.21	90.97	80.84	69.71	82.28	83.35	79.05	85.01	17.01	20.68	11.70	9.33	14.68	20.77	26.01	11.85	10.03	17.17	15.93	
ELCFS (Ours)	95.37	87.52	93.37	94.50	92.69	84.13	71.88	83.94	85.51	81.37	87.03	11.36	17.10	10.83	7.24	11.63	18.65	19.36	11.17	8.91	14.52	13.07	

Table 2. Comparison of federated domain generalization results on prostate MRI segmentation.

Unseen Site	A	B	C	D	E	F	Average	A	B	C	D	E	F	Average
	Dice Coefficient (Dice) \uparrow						Hausdorff Distance (HD) \downarrow							
JiGen [3]	89.95	85.81	84.06	87.34	81.32	89.11	86.26	10.51	11.53	11.70	11.49	14.80	9.02	11.51
BigAug [60]	89.63	84.62	83.86	87.66	81.20	88.96	85.99	10.68	11.78	12.07	10.66	13.98	9.73	11.48
Epi-FCR [25]	89.72	85.39	84.97	86.55	80.63	89.76	86.17	10.60	12.31	12.29	12.00	15.68	8.81	11.95
RSC [17]	88.86	85.56	84.36	86.21	79.97	89.80	85.80	10.57	11.84	14.76	13.07	14.79	8.83	12.31
FedAvg [36]	89.02	84.48	84.11	86.30	80.38	89.15	85.57	11.64	12.01	14.86	11.80	14.90	9.30	12.42
ELCFS (Ours)	90.19	87.17	85.26	88.23	83.02	90.47	87.39	10.30	11.49	11.50	11.57	11.08	8.31	10.88

results with two cases from unseen domains for each task. It is observed that our method accurately segments the structure and delineates the boundary in images of unknown distributions, whereas other methods sometimes fail to do so.

4.4. Ablation Analysis of Our Method

We conduct ablation studies to investigate four key questions regarding our ELCFS: **1)** the contribution of each component to our model performance, **2)** the benefit of the interpolation operation and the choice of λ , **3)** how the semantic feature space around the boundary region is influenced by our method, and **4)** how the numbers of participating clients affect the performance of our method.

Contribution of each component: We first validate the effect of the two key components in our method, i.e. continuous frequency space interpolation (**CFSI**) and Boundary-oriented Episodic Learning (**BEL**), by removing them respectively from our method to observe the model performance. As shown in Fig. 5, removing either part will lead to decrease on the generalization performance in different unseen domain settings for the two tasks. This is reasonable and reflects how the two components play complementary roles to the performance of our method, i.e., the generated distributions from CFSI lays foundation for the learning of BEL, and the BEL inversely provides assurance to effectively exploit the generated distributions.

Importance of continuous interpolation in frequency space: To analyze the effect of continuous interpolation mechanism in ELCFS , we use t-SNE [34] to visualize the distribution of generated images in fundus image segmentation. As shown in Fig. 6 (a), the pink points denote the local data of a client, and other points denote the transformed data that are generated with amplitude spectrum from different clients. It appears that fixing λ (left) will lead to several dis-

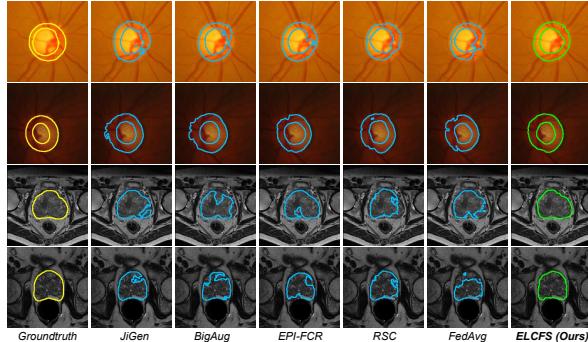


Figure 4. Qualitative comparison on the generalization results of different methods in fundus image segmentation (top two rows) and prostate MRI segmentation (bottom two rows).

tinct distributions, while the continuous interpolation mechanism (right) can smoothly bridge the distinct distributions to enrich the established multi-domain distributions. This promotes the local learning to attain domain-invariance in a dedicated dense distribution space.

We then analyze the effect of the choice of λ on our model performance, for which we conduct experiments with fixed values from 0.0 to 1.0 with a step size 0.2, and continuous sampling in range of [0.0, 0.5], [0.5, 1.0] and [0.0, 1.0]. As shown in Fig. 6 (b), compared with not transferring any distribution information (i.e., $\lambda = 0$), setting $\lambda > 0$ as a fixed value can always improve the model performance. Besides, the continuous sampling can further improve the performance and the sampling range of [0.0, 1.0] yields the best results, which reflects the benefits of continuous distribution space for domain generalization.

Discriminability at ambiguous boundary region: We plot the cosine distance between the boundary-related and background-related features, i.e., $\mathbb{E}[h_{i,bd} \odot h_{i,bg}]$, to ana-

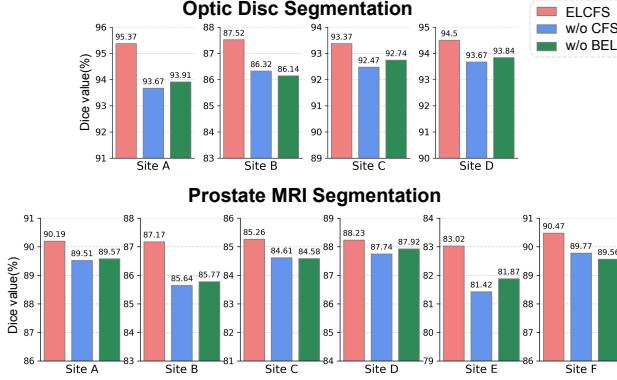


Figure 5. Ablation results to analyze the effect of the two components (i.e. CFSI and BEL) in our method.

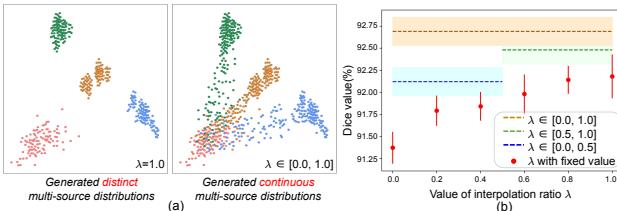


Figure 6. (a) Visualization of t-SNE [34] embedding for the original fundus images at a local client (pink points) and the corresponding transformed images with amplitude spectrum from different clients (green, yellow, and blue points); (b) Generalization performance on optic disc segmentation under different settings of interpolation ratio λ , with fixed value or continuous sampling from different ranges (with error bar from three independent runs).

lyze how the semantic feature space around the boundary region is influenced by our method. In Fig. 7 (a), the two green lines denote the growth of feature distance in our ELCFS and the FedAvg baseline respectively, for samples drawn from the training source domains. We can see that ELCFS yields a higher feature distance, indicating that the features of the boundary and the surrounding background region can be better separated in our method. For the two yellow lines, sample features are drawn from the unseen domains. As expected, the distance is not as high as in source domain, yet our method also presents a clearly higher margin than FedAvg. We also quantitatively analyze the effect of $\mathcal{L}_{boundary}$ on the model performance. As observed from Fig. 7 (b), removing this objective from the meta optimization leads to consistent performance drops on the generalization performance in different tasks.

Effect of participating client number: We further analyze how the generalization performance of our method and FedAvg will be affected when different numbers of hospitals participating in federated learning. Fig. 8 shows the results on prostate MRI segmentation, in which we present the generalization results on two unseen sites with the client number gradually increasing from 1 to $K - 1$. As expected, the models trained with single-source data cannot obtain

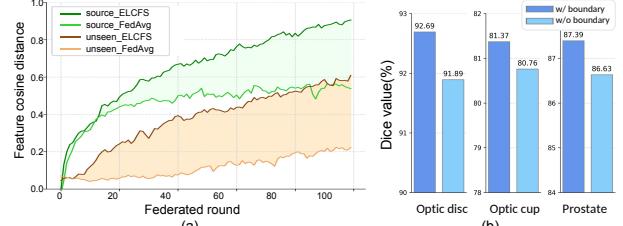


Figure 7. (a) Cosine distance between the boundary-related and background-related features; (b) Generalization performance of our method with or without the boundary-oriented meta objective.

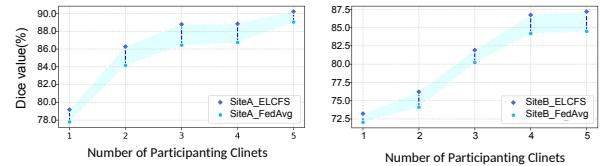


Figure 8. Curves of generalization performance on two unseen prostate datasets (i.e., site A and B) as the number of participating clients increases, using our proposed approach and FedAvg.

good results when deployed to unseen domains. The generalization performance increases when more clients participating in the federated training, which is reasonable as aggregating data from multiple sources can cover a more comprehensive data distribution. Particularly, our ELCFS consistently outperforms FedAvg on all generalization settings with different client numbers, demonstrating the stable efficacy of our method to leverage distributed data sources to enhance the generalizability of federated learning model.

5. Conclusion

We have proposed a novel problem setting of federated domain generalization, and presented a novel approach for it with continuous frequency space interpolation and a boundary-oriented episodic learning scheme. The superior efficacy of our method is demonstrated on two important medical image segmentation tasks. Our solution has opened a door in federated learning to enable local client access multi-source distributions without privacy leakage, which has great potential to address other problems encountered in FL, e.g., data heterogeneity. The proposed learning scheme for encouraging boundary delineation is also generally extendable to other segmentation problems.

6. Acknowledgement

This work was supported by Key-Area Research and Development Program of Guangdong Province, China (2020B010165004); National Natural Science Foundation of China with Project No. U1813204; Hong Kong Innovation and Technology Fund (Project No. ITS/311/18FP and GHP/110/19SZ).

References

- [1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, pages 998–1008, 2018. 2
- [2] N Bloch, A Madabhushi, H Huisman, J Freymann, J Kirby, M Grauer, A Enquobahrie, C Jaffe, L Clarke, and K Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370, 2015. 6, 12
- [3] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 1, 3, 6, 7, 12
- [4] Qi Chang, Hui Qu, Yikai Zhang, Mert Sabuncu, Chao Chen, Tong Zhang, and Dimitris N Metaxas. Synthetic learning: Learn from distributed asynchronous discriminator gan without sharing medical image data. In *CVPR*, pages 13856–13866, 2020. 1, 2
- [5] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. *arXiv preprint arXiv:2008.12839*, 2020. 2
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 5
- [7] Sanket S Dhruva, Joseph S Ross, Joseph G Akar, Brittany Caldwell, Karla Childers, Wing Chow, Laura Ciaccio, Paul Coplan, Jun Dong, Hayley J Dykhoff, et al. Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. *npj Digital Medicine*, 3(1):1–9, 2020. 1
- [8] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6450–6461, 2019. 1, 2
- [9] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. *arXiv preprint arXiv:2007.07645*, 2020. 2
- [10] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose 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)*, pages 1–6. IEEE, 2011. 6, 12
- [11] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 1, 2
- [12] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2477–2486, 2019. 2
- [13] Nathalie Guyader, Alan Chauvin, Carole Peyrin, Jeanny Héroult, and Christian Marendaz. Image phase or amplitude? rapid scene categorization is an amplitude-based process. *Comptes Rendus Biologies*, 327(4):313–318, 2004. 2
- [14] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefer, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020. 2
- [15] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. *arXiv preprint arXiv:2003.08082*, 2020. 2
- [16] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*, 2017. 2
- [17] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *ECCV*, 2020. 3, 6, 7, 12
- [18] Georgios A Kaassis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, pages 1–7, 2020. 1, 2
- [19] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 3
- [20] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 1, 2
- [21] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 6, 12
- [22] Daiqing Li, Amlan Kar, Nishant Ravikumar, Alejandro F Frangi, and Sanja Fidler. Federated simulation for medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 159–168. Springer, 2020. 2
- [23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 3
- [24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. *arXiv preprint arXiv:1710.03463*, 2017. 1, 2
- [25] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1446–1455, 2019. 1, 3, 6, 7, 12
- [26] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 1, 2

- [27] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 133–141. Springer, 2019. [2](#)
- [28] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *arXiv preprint arXiv:2001.05647*, 2020. [2](#)
- [29] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. [1, 2](#)
- [30] Yiyi Li, Yongxin Yang, Wei Zhou, and Timothy M. Hospedales. Feature-critic networks for heterogeneous domain generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3915–3924. PMLR, 2019. [2](#)
- [31] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. [5, 6, 12](#)
- [32] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 475–485. Springer, 2020. [2](#)
- [33] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE transactions on medical imaging*, 39(9):2713–2724, 2020. [6](#)
- [34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. [7, 8](#)
- [35] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, pages 11749–11756, 2020. [3](#)
- [36] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017. [2, 3, 6, 7, 12](#)
- [37] Saeid Motian, Marco Piccirilli, Donald A Adjeroh, and Gi-anfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725, 2017. [1, 2](#)
- [38] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant fea-
ture representation. In *International Conference on Machine Learning*, pages 10–18, 2013. [2](#)
- [39] Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, pages 80–111. Springer, 1981. [3](#)
- [40] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. [5, 6, 12](#)
- [41] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *arXiv preprint arXiv:1911.02054*, 2019. [2](#)
- [42] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982. [2](#)
- [43] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. [1, 2](#)
- [44] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *arXiv preprint arXiv:2003.08119*, 2020. [1](#)
- [45] Holger R Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C Bizzo, et al. Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 181–191. Springer, 2020. [2](#)
- [46] Hermann Schomberg and Jan Timmer. The gridding method for image reconstruction by fourier transformation. *IEEE TMI*, 14(3):596–607, 1995. [4](#)
- [47] Seonguk Seo, Yumin Suh, Dongwan Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. *ECCV*, 2020. [2](#)
- [48] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. [3](#)
- [49] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion Workshop, MICCAI*, pages 92–104. Springer, 2018. [1, 2](#)
- [50] Smadar Shilo, Hagai Rossman, and Eran Segal. Axes of a revolution: challenges and promises of big data in healthcare. *Nature Medicine*, 26(1):29–38, 2020. [1](#)
- [51] Santiago Silva, Boris A Gutman, Eduardo Romero, Paul M Thompson, Andre Altmann, and Marco Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *ISBI*, pages 270–274. IEEE, 2019. [2](#)

- [52] Jayanthi Sivaswamy, S Krishnadas, Arunava Chakravarty, G Joshi, A Syed Tabish, et al. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1):1004, 2015. [6](#), [12](#)
- [53] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, pages 5334–5344, 2018. [3](#)
- [54] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pages 159–176. Springer, 2020. [3](#)
- [55] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging*, 39(12):4237–4248, 2020. [2](#)
- [56] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. [2](#)
- [57] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *CVPR*, 2020. [2](#)
- [58] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2100–2110, 2019. [2](#)
- [59] Sergey Zakharov, Wadim Kehl, and Slobodan Ilic. Deceptionnet: Network-driven domain randomization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 532–541, 2019. [2](#)
- [60] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 2020. [3](#), [6](#), [7](#), [12](#)
- [61] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, pages 561–578. Springer, 2020. [3](#)

7. Supplementary Material

7.1. Datasets Details:

The retinal fundus images adopted in the experiments are collected from four different clinical centers out of three public datasets. Among these data, samples of sites A are from Drishti-GS [52] dataset; samples of site B are from RIM-ONE-r3 [10] dataset; samples of site C, D are from REFUGE [40] dataset. Note that the REFUGE dataset includes two different data sources, so we decompose them in our federated learning setting. Among the six data sources in the prostate MRI segmentation task, samples of Site A, B are from NIC-ISBI13 [2] datasets; samples of Site C are from I2CVB [21] datasets; and samples of Site D, E, F are from PROMISE12 [31] dataset. Similarly, since the NIC-ISBI13 and PROMISE12 contain data from multiple data sources, we separate them and consider each data source as an individual client in the federated scenario. Details of the scanners and imaging protocols of these data are illustrated in Table 3 and Table 4 respectively.

Table 3. Details of the scanning protocols for different data sources in fundus image segmentation.

Task	Dataset	Manufactor
Fundus Image Segmentation	Site A [52]	(Aravind eye hospital)
	Site B [10]	Nidek AFC-210
	Site C [40]	Zeiss Visucam 500
	Site D [40]	Canon CR-2

Table 4. Details of the scanning protocols for different data sources in prostate MRI segmentation.

Task	Dataset	Manufactor	Field strength(T)	Endorectal Coil
Prostate MRI Segmentation	Site A [2]	Siemens	3	Surface
	Site B [2]	Philips	1.5	Endorectal
	Site C [21]	Siemens	3	No
	Site D [31]	Siemens	1.5 and 3	No
	Site E [31]	GE	3	Endorectal
	Site F [31]	Siemens	1.5	Endorectal

7.2. Statistical Analysis

We conduct paired t-test between our approach and different comparison methods to analyze whether the performance improvement of our approach is significant. We adopt Dice as the evaluation measurement and set the significance level as 0.05. For each method, the statistical tests are conducted by jointly considering the prediction results of each unseen site setting on overall generalization performance. The results are listed in Table 5. It is observed that all paired t-test results present p -value smaller than 0.05, demonstrating that our improvements over these state-of-the-art domain generalization methods are significant.

Table 5. P-value for statistical analysis between our approach and different comparison methods on overall Dice score.

	JiGen [3]	BigAug [60]	Epi-FCR [25]	RSC [17]	FedAvg [36]
Optic disc	3.6e-20	7.6e-22	3.5e-12	2.1e-9	1.2e-8
Optic cup	1.1e-16	0.0026	1.6e-7	0.0003	2.0e-21
Prostate	0.0004	2.3e-7	9.2e-5	5.2e-8	2.9e-8

7.3. Visualization of Transformed Data

We visualize the appearances of transformed images under different interpolation ratio λ for the two tasks. As shown in Fig. 9, the appearance of local source image is indeed gradually transformed to the style (i.e. distribution) of target image of other clients as we increase the interpolation ratio from 0 to 1, while the semantic content of the image is unchanged. Such continuous interpolation mechanism helps to enrich the multi-source distributions to a dedicated dense distribution space, hence benefits the model to gain domain-invariance in a more continuous latent space to improve the generalizability.

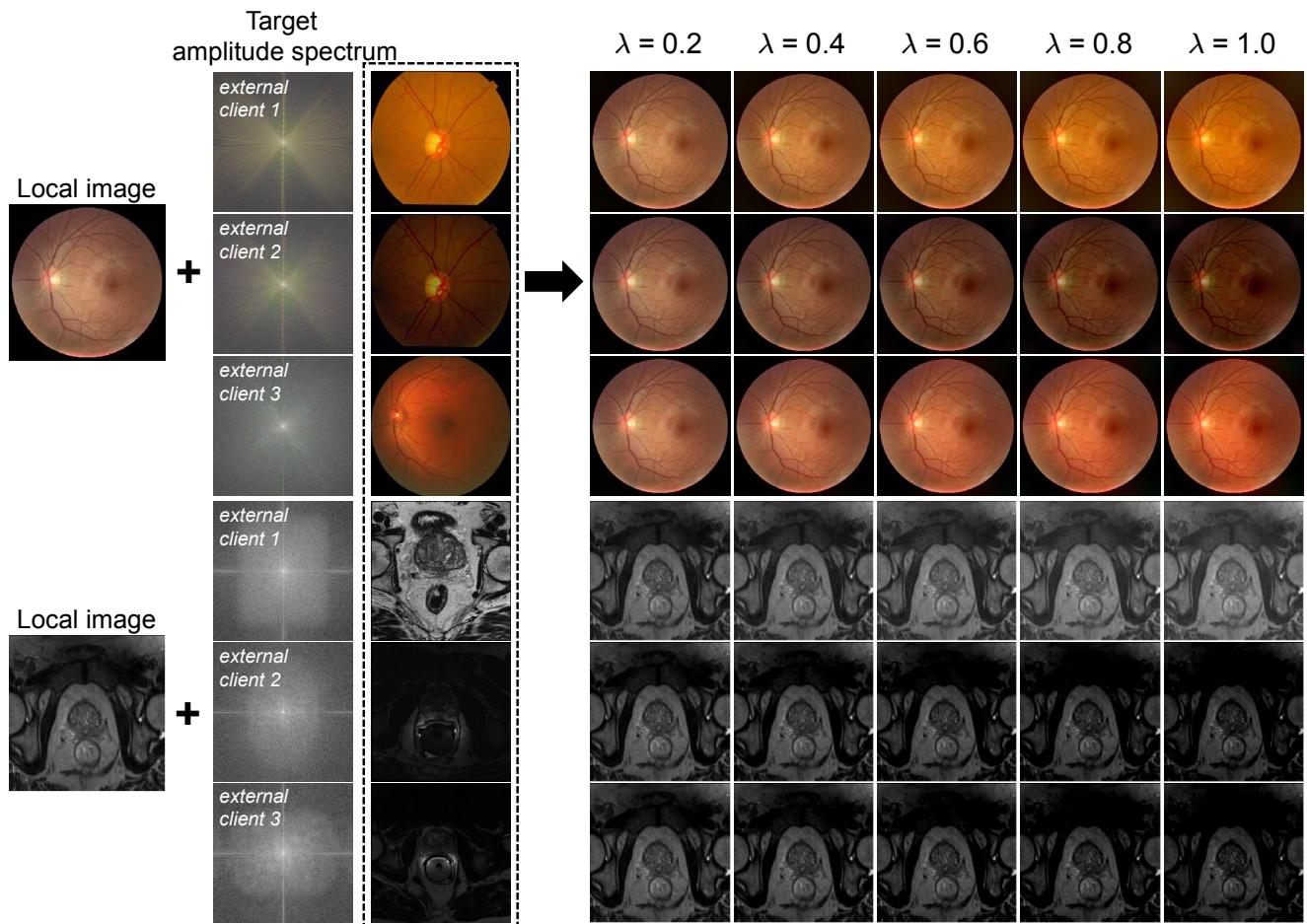


Figure 9. Visualization of transformed images under different interpolation ratio λ .