Vicinal Feature Statistics Augmentation for Federated 3D Medical Volume Segmentation

Yongsong Huang[1,2], Wanqing Xie[1,3], Mingzhen Li[1,4], Mingmei Cheng[1,2], Jinzhou Wu[1], Weixiao Wang[1], Jane You[5], and Xiaofeng Liu[1(✉)]

[1] Harvard Medical School, Harvard University, Boston, MA 02114, USA
xliu61@mgh.harvard.edu
[2] Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan
[3] Department of Intelligent Medical Engineering, School of Biomedical Engineering, Anhui Medical University, Hefei 230032, China
[4] Department of Mathematics, Washington University in St. Louis, St. Louis, MO 63130, USA
[5] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Abstract. Federated learning (FL) enables multiple client medical institutes collaboratively train a deep learning (DL) model with privacy protection. However, the performance of FL can be constrained by the limited availability of labeled data in small institutes and the heterogeneous (i.e., non-i.i.d.) data distribution across institutes. Though data augmentation has been a proven technique to boost the generalization capabilities of conventional centralized DL as a "free lunch", its application in FL is largely underexplored. Notably, constrained by costly labeling, 3D medical segmentation generally relies on data augmentation. In this work, we aim to develop a vicinal feature-level data augmentation (VFDA) scheme to efficiently alleviate the local feature shift and facilitate collaborative training for privacy-aware FL segmentation. We take both the inner- and inter-institute divergence into consideration, without the need for cross-institute transfer of raw data or their mixup. Specifically, we exploit the batch-wise feature statistics (e.g., mean and standard deviation) in each institute to abstractly represent the discrepancy of data, and model each feature statistic probabilistically via a Gaussian prototype, with the mean corresponding to the original statistic and the variance quantifying the augmentation scope. From the vicinal risk minimization perspective, novel feature statistics can be drawn from the Gaussian distribution to fulfill augmentation. The variance is explicitly derived by the data bias in each individual institute and the underlying feature statistics characterized by all participating institutes. The added-on VFDA consistently yielded marked improvements over six advanced FL methods on both 3D brain tumor and cardiac segmentation.

1   Introduction

Federated learning (FL) [25] for medical image analysis, which follows a privacy-aware decentralized paradigm to learn a global model on several local medical institutes, has recently been the center of much attention. In practice, however, the participating institutes may have diverse data collection schemes, which can lead to inefficient global collaboration [17,25] in two ways. First, some of the small institutes may have limited training data to support effective local updating. This is especially evident in the medical segmentation task, which is often constrained by the availability of costly labeled training data [7]. Second, the data collected from different institutes with heterogeneous vendors, doses, and populations can result in biased gradient uploading to hinder the convergence of the federal model. Though in conventional centralized learning, the widely used data augmentation [27] can be a straightforward and unified solution to the challenges of limited sample and poor generalized data distribution [37], it remains largely underexplored in FL segmentation.

It is of great importance to explore the proper global data augmentation scheme under the strict privacy restriction of FL, in which each client cannot access the data across institutes [25]. Simply utilizing the data augmentation methods in centralized learning without injecting global information is sub-optimal, inheriting the institute's bias. Early attempts [13,15,28] require accessing a global dataset to achieve balanced training, which is restricted in standard FL settings. Astraea [6] addresses this issue with several mediators between global servers and institutes. However, since each mediator group is a set of institutes, privacy among them is not protected. On the other hand, FedMix [34] and XORMix [26] propose to transmit the averaged images across institutes, which still take risk of privacy breaches and are inherently weak at constructing semantic transform with image-level MixUp [18]. In addition, the above methods focus on classification and do not apply to segmentation–an important medical image analysis task with great demand for data augmentation [7].

In this work, we propose to take both the inner- and inter-institute divergence into consideration without the need for any cross-institute transmission of raw data or their MixUp [26,34]. To mitigate the aforementioned limitations, we resort to the feature-level vicinal risk minimization (VRM) [4] to expand an example point in the feature space to a probabilistically modeled flexible distribution, e.g., a conceptually simple Gaussian prototype, centered at the original point, with the variation reflecting the local and global data divergence. Novel feature statistics can be drawn from the Gaussian prototype distribution to fulfill augmentation. Moreover, inspired by the previous image style generation and domain adaptation works [3,16,20], we propose to exploit the batch-wise feature statistics (e.g., mean and standard deviation) in each institute to abstractly represent the discrepancy of data, which are shared among institutes without any raw data transmission.

With our VFDA framework, the core issue is to properly associate the variance with the data bias in each individual institute and the underlying feature statistics characterized by all participating institutes. For effective augmentation,

a proper variance is determined based on variances of feature statistics within each institute, regulated by the global variance of feature statistics characterized by all participating institutes. The augmentation in VFDA allows the local model to be trained over samples drawn from diverse feature distributions, alleviating local distribution shifts and facilitating institute-invariant representation learning, eventually leading to a better global model.

The main contributions of this work can be summarized as follows:

- To our knowledge, this is the first attempt to investigate an efficient data augmentation scheme in FL segmentation, which is especially pertinent for 3D medical data analysis.
- We propose a vicinal feature-level data augmentation (VFDA) scheme to model each feature statistic probabilistically via a conceptually simple Gaussian prototype and use the statistics drawn from the Gaussian distribution to implement augmentation. The transmission of raw images or their MixUp is not needed, meeting requirements for strict privacy protection.
- Both local and global divergence are taken into account in the abstractly represented batch feature statistics to quantify the augmentation scope.
- We evaluated our VFDA scheme on 3D brain tumor and cardiac anatomical segmentation tasks with 6 advanced FL methods to demonstrate its general efficacy and superiority.

2   Methodology

In FL segmentation, we are given N local client institutes, in which the n-th institute has M^n pairs of input volume $x_m^n \in \mathcal{X}$ and the corresponding segmentation label $y_m^n \in \mathcal{Y}$. We are expected to learn a global segmentation model $f_g(w_g) : \mathcal{X} \to \mathcal{Y}$ parameterized by w_g across all institutes, generally enforced by the global empirical risk minimization (ERM) objectives:

$$\mathcal{L}^g_{ERM}(w_g) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{(x_m^n, y_m^n) \sim \mathcal{P}^n}[\mathcal{L}^n_{ERM}(x_m^N, y_m^N; w_n)], \qquad (1)$$

where $\mathcal{P}^n$ is the underlying distribution of institute n (i.e., $\{x_m^n, y_m^n\}_{m=1}^{M^n} \sim \mathcal{P}^n$), and $\mathcal{L}^n_{ERM}$ is the institutes-wise empirical risk with local network parameter w_n. Specifically, in the segmentation task, $\mathcal{L}^n_{ERM}$ typically follows the empirical form of voxel-wise cross-entropy or Dice loss. With a privacy-aware decentralized setting, directly calculating $\mathcal{L}^g_{ERM}(w_g)$ with all data across institutes as conventional centralized training is infeasible. In contrast, FL [21,25] relies on local training of models $f_n(w_n), n \in \{1, \cdots, N\}$ in each institute in parallel with the local data only. In each round, the trained local models are aggregated to a global model $f_g(w_g)$, which is then further distributed to each institute for the next round of local training. Therefore, the local training objective in FL is equivalent to empirically approximating the local distribution $\mathcal{P}^n$ by a finite M^n number of samples, i.e., $\mathcal{P}^n_{ERM}(x, y) = \frac{1}{M^n} \sum_{m=1}^{M^n} \delta(x = x_m^n, y = y_m^n)$, where $\delta(x = x_m^n, y = y_m^n)$ is a Dirac delta distribution with a point mass at (x_m^n, y_m^n).
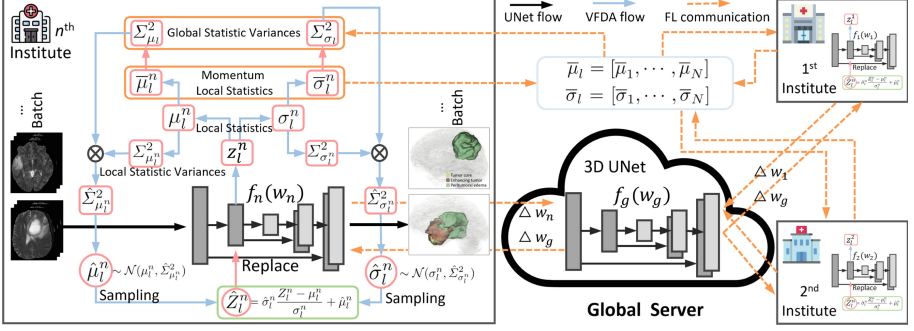
Fig. 1. Illustration of our VFDA over an encoder layer in the classical FL 3D segmentation framework with heterogeneous clients.

Although the ERM objectives have been widely adopted in deep FL frameworks as the training scheme, the underlying assumption is that the empirical local distributions $\mathcal{P}^n_{ERM}$ are homologous to the underlying global distribution $\mathcal{P}^g$, which is unrealistic in actual clinical scenarios [16,24]. In practice, there can be a significant performance drop since each $\mathcal{P}^n_{ERM}$ exhibits diverse data drifts with respect to $\mathcal{P}^n$ and $\mathcal{P}^g$, which lead to inconsistency of local and global empirical objectives and difficulties of generalization to testing distribution [1,9]. In addition, each local model can be data starved, given only access to institute samples.

2.1 Label Consistent Vicinal Feature Distribution Extrapolation

Based on the above concerns, we propose to expand the Dirac delta distribution inherent in $\mathcal{P}^n_{ERM}(x,y)$ to a more expressive one to approximate the true distribution following the idea of VRM [4,36]. Therefore, we are able to examine the vicinal region of each sample (x^n_m, y^n_m) for infinite data augmentation. Specifically, we can define a institute-wise vicinity distribution $\mathcal{V}^n(\hat{x}^n_m, \hat{y}^n_m | x^n_m, y^n_m)$ and apply it to each sample (x^n_m, y^n_m) to generate numerous pseudo samples $(\hat{x}^n_m, \hat{y}^n_m)$ and support the local institute training. More formally, we construct the vicinal local distribution as $\mathcal{P}^n_{VRM} = \frac{1}{M^n} \sum_{m=1}^{M^n} \mathcal{V}^n(\hat{x}^n_m, \hat{y}^n_m | x^n_m, y^n_m)$, and expect a better mimic of $\mathcal{P}^n$ with a proper $\mathcal{V}^n(\cdot)$.

There are numerous successful candidates of $\mathcal{V}^n(\cdot)$ in conventional centralized learning, e.g., MixUp [36] and CutMix [35]. Though simply utilizing MixUp [36] or CutMix locally can potentially improve the performance (as shown in experiments), it can be sub-optimal as there are no global cues injected. In these cases, $\mathcal{P}^n_{VRM}$ gives a better mimic of the true local distribution $\mathcal{P}^n$ rather than the global distribution $\mathcal{P}^g$.

Instead of the raw voxel level MixUp [35,36], VFDA estimates a vicinity distribution $\mathcal{V}^n_l$ at each encoder layer l to augment its batch-wise intermediate feature Z^n_l in n-th institute for more flexible expansion [18]. Of note, $Z^n_l \in \mathbb{R}^{B \times C \times H \times W \times S}$ denotes the intermediate feature representation of B mini-batch volumes, with height H, width W, slices S, and channels C. In addition,

to achieve consistency training [32], we define $\mathcal{V}_l^n$ to be label consistent, i.e., $\mathcal{V}_l^n(\hat{Z}_l^n, \hat{Y}^n | Z_l^n, Y^n) = \mathcal{V}_l^n(\hat{Z}_l^n | Z_l^n)\delta(\hat{Y}^n = Y^n)$, which only extrapolate the latent feature Z_l^n while maintaining the consistency of the label $Y^n \in \mathbb{R}^{B \times H \times W \times S}$. Then, a key challenge is adaptively configuring $\mathcal{V}_l^n(\hat{Z}_l^n | Z_l^n)$ based on the local and global data divergence.

2.2   Probabilistic Modeling of Feature Statistics

As opposed to explicitly modeling $\mathcal{V}_l^n(\hat{Z}_l^n | Z_l^n)$, VFDA resorts to implicit feature augmentation by exploiting the batch-wise feature statistics in each institute to abstractly represent the discrepancy of input data, and model each feature statistic probabilistically via a Gaussian prototype. Specifically, we utilize the channel-wise feature statistics of mean μ_l^n and standard deviation σ_l^n. For Z_l^n, its channel-wise statistics of $\mu_l^n \in \mathbb{R}^{B \times C}$ and $\sigma_l^n \in \mathbb{R}^{B \times C}$ can be formulated as:

$$\mu_l^n = \frac{1}{H \times W \times S} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{s=1}^{S} Z_l^n; \quad \sigma_l^n = \sqrt{\frac{1}{H \times W \times S} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{s=1}^{S} (Z_l^n - \mu_l^n)}. \tag{2}$$

Recent works [3,19,20,30] demonstrated that these low-order batch statistics are domain-specific, owing to the divergence of feature representations. As the abstraction of latent features, the feature statistics among local institutes will also exhibit inconsistency and follow shifts from the statistics of the true distribution.

We propose to capture such shifts via probabilistic modeling. We hypothesize that each feature statistic follows a multi-variate Gaussian prototype, i.e., $\mu_l^n \sim \mathcal{N}(\mu_l^n, \hat{\Sigma}_{\mu_l^n}^2)$ and $\sigma_l^n \sim \mathcal{N}(\sigma_l^n, \hat{\Sigma}_{\sigma_l^n}^2)$, where each Gaussian prototype's center corresponds to the original statistic, and the variance is an estimation of the potential feature statistic shift from the true distribution.

2.3   Local and Global Statistic Variances Quantification

Determining an appropriate variant range is challenging since each institute has only access to the data itself but has no sense of its statistical biases. Recent research [29,31] has shown that deep feature space contains many semantic directions, and feature variances provide a reasonable measurement of potential meaningful semantic changes along the directions. This motivates us to estimate the variance of the Gaussian prototype from the variance of feature statistics.

Local Statistic Variances. In each institute, we compute local variances of feature statistics based on the information within each mini-batch:

$$\Sigma_{\mu_l^n}^2 = \frac{1}{B} \sum_{b=1}^{B} (\mu_l^n - \mathbb{E}[\mu_l^n])^2 \in \mathbb{R}^C; \quad \Sigma_{\sigma_l^n}^2 = \frac{1}{B} \sum_{b=1}^{B} (\sigma_l^n - \mathbb{E}[\sigma_l^n])^2 \in \mathbb{R}^C, \tag{3}$$

where $\Sigma_{\mu_l^n}^2$ and $\Sigma_{\sigma_l^n}^2$ denote the variance of feature mean μ_l^n and standard deviation σ_l^n that are specific to each institute. Each value in $\Sigma_{\mu_l^n}^2$ and $\Sigma_{\sigma_l^n}^2$ is the

variance of feature statistics in a particular channel, and its magnitude captures the potential change in that particular channel at that specific institute.

Global Statistic Variances. The institute-specific statistic variances are solely computed based on the data in each institute and are thus likely biased due to the bias in the local dataset. To resolve this, we further estimate the variances of the institute-sharing feature statistics, taking information from all institutes into account. Particularly, we propose a momentum version of feature statistics for each institute, which is updated online with an exponential momentum decay (EMD) strategy:

$$\overline{\mu}_l^n \leftarrow (1-\eta)\frac{1}{B}\sum_{b=1}^{B}\mu_l^n + \eta\overline{\mu}_l^n \in \mathbb{R}^C; \quad \overline{\sigma}_l^n \leftarrow (1-\eta)\frac{1}{B}\sum_{b=1}^{B}\sigma_l^n + \eta\overline{\sigma}_l^n \in \mathbb{R}^C, \quad (4)$$

where the momentum factor $\eta = \eta^0\exp(-r)$ follows an exponential decay over round r. Notably, $\overline{\mu}_l^n$ and $\overline{\sigma}_l^n$ are the momentum updated feature statistics of encoder layer l in institute n. η^0 is empirically initialized to 10.

In each communication round, these accumulated local feature statistics are sent to the server along with model parameters. Let $\overline{\mu}_l = [\overline{\mu}_1, \cdots, \overline{\mu}_N] \in \mathbb{R}^{N\times C}$ and $\overline{\sigma}_l = [\overline{\sigma}_1, \cdots, \overline{\sigma}_N] \in \mathbb{R}^{N\times C}$ denote the collections of accumulated feature statistics of all institutes, then the global, institute sharing statistic variances are calculated as:

$$\Sigma_{\mu_l}^2 = \frac{1}{N}\sum_{n=1}^{N}(\overline{\mu}_l^n - \mathbb{E}[\overline{\mu}_l])^2 \in \mathbb{R}^C; \quad \Sigma_{\sigma_l}^2 = \frac{1}{N}\sum_{n=1}^{N}(\overline{\sigma}_l^n - \mathbb{E}[\overline{\sigma}_l])^2 \in \mathbb{R}^C. \quad (5)$$

Along with the aggregated model parameters as in classical FL methods, e.g., FedAvg [21], these variances are distributed back to each institute to inform a global estimation of feature statistic variances. Note that $\Sigma_{\mu_l}^2$ and $\Sigma_{\sigma_l}^2$ are shared by all participating institutes.

Institute sharing estimations $\Sigma_{\mu_l}^2$ and $\Sigma_{\sigma_l}^2$ provide a quantification of distribution divergence among institutes, and larger values imply potentials of more significant changes of the corresponding channels in the true feature statistic space. Therefore, for each institute, we weight the institute-specific statistic variances $\Sigma_{\mu_l^n}^2$, $\Sigma_{\sigma_l^n}^2$ with $\Sigma_{\mu_l}^2$, $\Sigma_{\sigma_l}^2$, so that each institute has a sense of such global divergence, i.e., $\hat{\Sigma}_{\mu_l^n}^2 = \Sigma_{\mu_l}^2\Sigma_{\mu_l^n}^2$ and $\hat{\Sigma}_{\sigma_l^n}^2 = \Sigma_{\sigma_l}^2\Sigma_{\sigma_l^n}^2$.

2.4   Implementation of Vicinal Feature-Level Data Augmentation

After establishing the Gaussian prototype, we calculate novel feature $\hat{Z}_l^n$ in the vicinity of Z_l^n as follows:

$$\hat{Z}_l^n = \hat{\sigma}_l^n\frac{Z_l^n - \mu_l^n}{\sigma_l^n} + \hat{\mu}_l^n; \quad \hat{\mu}_l^n \sim \mathcal{N}(\mu_l^n, \hat{\Sigma}_{\mu_l^n}^2), \quad \hat{\sigma}_l^n \sim \mathcal{N}(\sigma_l^n, \hat{\Sigma}_{\sigma_l^n}^2), \quad (6)$$

where Z_l^n is first normalized with its original statistics by $\frac{Z_l^n - \mu_l^n}{\sigma_l^n}$, and further scaled with novel statistics $\hat{\mu}_l^n$ and $\hat{\sigma}_l^n$ that are randomly sampled from the

Table 1. Comparisons of different FL methods with or without our VFDA framework in FeTS2021, and the ablation study of VFDA modules.

Method	Additional augmentation	Dice	Score	[%] ↑	
		ET	TC	WT	Mean
FedAvg [21]	–	75.63	62.01	76.55	71.14
FedAvg [21]	MixUp	75.80	62.74	76.95	71.83
FedAvg [21]	VFDA (Ours)	**76.59**	**64.10**	**77.86**	**72.85**
FedAvg [21]	VFDA w/o EMD	76.08	63.47	77.29	72.28
FedAvg [21]	VFDA w/o Global Statistic Variances	75.92	63.15	77.03	72.03
FedNorm [33]	–	75.60	63.76	76.98	72.11
FedNorm [33]	MixUp	75.82	63.96	77.05	72.28
FedNorm [33]	VFDA (Ours)	**76.49**	**64.82**	**78.17**	**73.16**
FedNorm [33]	VFDA w/o EMD	76.14	64.55	77.83	72.84
FedNorm [33]	VFDA w/o Global Statistic Variances	76.01	64.34	77.50	72.62

corresponding Gaussian distribution [12]. To make the sampling differentiable, we apply the re-parameterization trick [10]:

$$\hat{\mu}_l^n = \mu_l^n + \epsilon_\mu \hat{\Sigma}_{\mu_l^n}^2, \quad \hat{\sigma}_l^n = \sigma_l^n + \epsilon_\sigma \hat{\Sigma}_{\sigma_l^n}^2), \tag{7}$$

where $\epsilon_\mu \sim \mathcal{N}(0,1)$ and $\epsilon_\sigma \sim \mathcal{N}(0,1)$ follow normal distribution.

The proposed VFDA in Eq. (6) works in a plug-and-play fashion, which can be inserted at arbitrary positions of the model to facilitate latent semantic augmentation. In our implementation, we add a VFDA after each encoder layer of UNet. Of note, we explore the batch-wise feature statistics for vicinal expansion while not relying on the network with batch normalization layer [14,16]. Our FDA can be widely generalized to modern deep learning models with mini-batch training. During testing, no augmentation is performed.

3   Experiments and Results

To show the effectiveness of our framework, we experimented on both 3D federated brain tumor and cardiac anatomical segmentation tasks and added VFDA to numerous advanced FL segmentation models, e.g., FedAvg [21], FedProx [8], FedBN [14], FedNorm [33], PRRF [5], and FedCRLD [23]. Of note, our VFDA can be directly applied to 2D segmentation by setting $S = 0$.

We implemented all modules on a server with an NVIDIA A100 GPU and used the PyTorch toolbox. For the evaluation metrics, we employed the widely accepted Dice similarity coefficient (DSC), which measures the overlap between the predicted segmentation mask and the label.

3.1   Federated Brain Tumor Segmentation

The Federated Tumor Segmentation (FeTS) 2021 Challenge Task-1 [22] incorporates the magnetic resonance imaging (MRI) volumes and segmentation label of
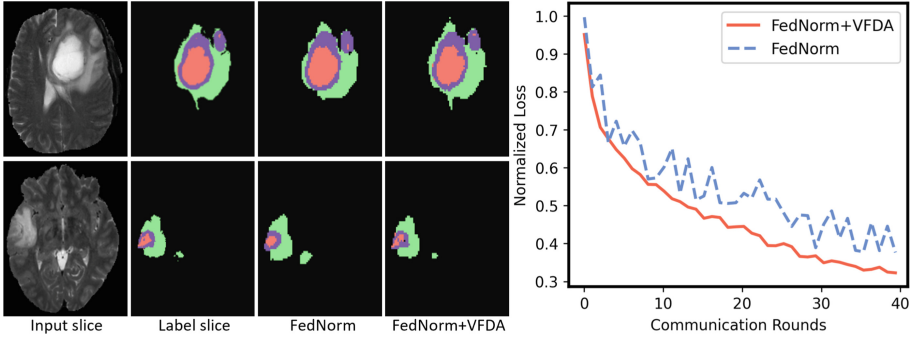
Fig. 2. Left: The qualitative comparisons of 2 slices from two subjects with FedNorm with or without our VFDA. CT: red; ET: purple; WT: red+purple+green; background: black (best viewed in color). Right: the normalized training loss of FedNorm with or without our VFDA.

three brain tumor structures, i.e., tumor core (TC), enhanced tumor (ET) and whole tumor (WT), from 341 subjects. For FL, we followed the standard dataset partition-2 to separate 22 local clients based on their original institutions and further split the large institutes into subsets according to the tumor sizes, which are subject-independent. MRI scans can be highly heterogeneous among participating clients in the FeTS challenge as various scanners and image protocols were employed. The size-based subset also involves the divergence of different tumor grades. With the fine-grained partition in FeTS partition-2, the subjects in each local client are limited.

According to the standard evaluation protocol, the segmentation model of ResNet-based 3D UNet is fixed for all participants [22]. We adopted the successful solutions in FeTS, i.e., FedNorm [33] and FedAvg [21] with different model aggregation schemes as our baselines. For fair comparisons, we followed the detailed setting of federated aggregation with tensor normalization (FedNorm) [33] based on the open-fl framework. All of the experiments were conducted under a fixed train validation split and random seed to make our results convincing and deterministic. Specifically, in each communication round, we performed one epoch for local client training and initialized the learning rate to 5e-4 with a polynomial decaying factor of 0.9 consistently. For all methods, we added the vanilla data augmentation of rotation, scaling, elastic deformation, brightness, and aggressiveness adjustment as in [33].

The quantitative evaluation results with respect to Dice score are presented in Table 1. Simply adding the MixUp augmentation improved the performance of both FedAvg and FedNorm significantly. By taking a more flexible feature-level expansion based on the local and global divergence described by the abstract feature statistic variances, our VFDA can achieve remarkable results. For the ablation study, we removed the global statistic variances or EMD module and denoted them as VFDA w/o EMD or VFDA w/o Global Statistic Variances,

Table 2. Comparisons of different FL methods with or without our VFDA framework in cardiac segmentation task, and the ablation study of VFDA modules.

Method	Additional augmentation	Dice	Score	[%] ↑				
		A	B	C	D	E	F	Mean
FedAvg [21]	–	85.84	85.39	89.08	79.77	84.42	18.36	73.81
FedAvg [21]	VFDA w/o Global	87.42	86.25	88.30	82.59	85.45	45.08	79.18
FedAvg [21]	VFDA (Ours)	**88.72**	**87.53**	**89.91**	**84.15**	**86.05**	**52.47**	**81.4**
FedProx [8]	–	86.70	84.41	88.81	82.85	83.66	27.26	75.62
FedProx [8]	VFDA w/o Global	87.65	86.40	89.28	84.91	85.13	49.72	80.51
FedProx [8]	VFDA (Ours)	**88.23**	**87.13**	**89.95**	**84.86**	**85.60**	**52.81**	**81.43**
FedBN [14]	–	86.98	85.87	89.58	81.91	84.73	27.49	76.09
FedBN [14]	VFDA w/o Global	88.05	86.85	89.74	83.62	85.82	45.65	79.96
FedBN [14]	VFDA (Ours)	**88.62**	**87.28**	**90.20**	**85.07**	**86.15**	**48.30**	**80.94**
PRRF [5]	–	87.04	86.11	88.05	84.65	83.85	53.09	80.47
PRRF [5]	VFDA w/o Global	87.86	87.03	89.65	85.12	85.44	61.05	82.69
PRRF [5]	VFDA (Ours)	**88.92**	**88.75**	**90.26**	**86.07**	**86.29**	**62.38**	**83.79**
FedCRLD [23]	–	88.06	87.28	90.88	86.96	86.40	76.15	85.96
FedCRLD [23]	VFDA w/o Global	89.02	88.34	91.24	87.35	87.23	77.80	86.83
FedCRLD [23]	VFDA (Ours)	**89.74**	**89.05**	**92.11**	**87.83**	**87.62**	**78.53**	**87.48**

respectively. Their inferior performance compared to VFDA demonstrates the contribution of global statistic variances or EMD training.

In Fig. 2, some example slices are shown, in which the VFDA achieves more accurate delineations compared to FedNorm. In addition, adding VFDA contributes to a stabler continuous optimization process than FedNorm.

3.2   Federated Cardiac Anatomical Segmentation

To further demonstrate the generalizability of VFDA, we also evaluated it on multi-center multi-sequence cardiac MRI segmentation as in [23]. Specifically, a real-world FL task is constructed using the publicly available M&M [2] and Emidec [11] datasets. Of note, M&M dataset incorporates the cine-MRI of the subjects from five centers in three countries and is scanned with four different scanner vendors, and the Emidec is a delayed enhancement (DE) MRI dataset. As in [23], the M&M dataset is split into five institutes/centers, i.e., client A-E, while Emidec is configured as the sixth institute, i.e., client F. There are notable appearance shifts across institute due to the diverge centers, devices, and contrast agents. We chose the subject-independent 7/1/2 split for each client institute and adopted the 3D UNet as the segmentation model backbone.

For fair comparisons, we adopted the vanilla augmentation of rotation, translation, scale, and mirror as [23]. Notably, the compared baselines of FedProx [8], FedBN [14], PRRF [5], and FedCRLD [23] are designed for non-IID cases, which explicitly target the data bias among local institutes using different techniques.
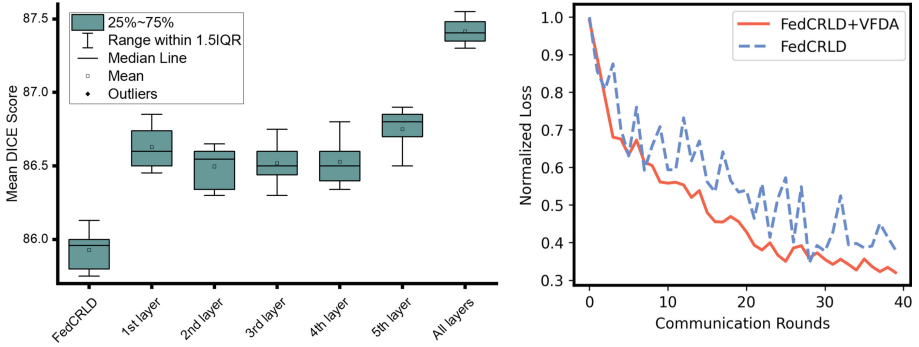
Fig. 3. Left: Sensitivity analysis of adding VFDA to each layer of the 3DUNet encoder in FedCRLD with five times random trails. Right: the normalized training loss of FedCRLD with or without our VFDA.

However, how to efficiently induce the global data divergence for FL debias is a long-lasting challenge. Our VFDA can be a unified and general solution to simply add on these methods to efficiently improve the segmentation performance, as shown in Table 2. In the ablation study, we also demonstrated that taking the global statistic variances into consideration is important for VFDA for all these baselines.

In Fig. 3 left, we investigated the effect of VFDA in different encoder layers. Of note, adding VFDA to the decoder layer does not improve the performance. In Fig. 3 right, the FedCRLD training in client F (i.e., Emidec with different DE-MRI) is relatively unstable, while adding VFDA smooths the optimization and leads to a lower loss.

4   Conclusion

In this work, we proposed a novel and efficient data augmentation methodology for federated learning in 3D medical volume segmentation, which suffers from the imbalance of clients in small institutes, and the inner- and inter-institute heterogeneous data shift. We resort to batch-wise feature statistics as an abstract quantification of the local and global statistic variances and utilize them probabilistically via a Gaussian prototype. We utilize the mean corresponding to the original statistic and the variance to define the proper augmentation scope in a label-preserving vicinal risk minimization framework on the feature space to expand the feature by simply sampling the Gaussian distribution with the re-parameterization trick. The experiments in both 3D brain tumor and cardiac anatomical structure segmentation FL tasks with six advanced FL methods consistently demonstrate its efficiency and generality. It has the potential to be widely adapted to different FL scenarios with low additional community costs.

References

1. Acar, D.A.E., Zhao, Y., Navarro, R.M., Mattina, M., Whatmough, P.N., Saligrama, V.: Federated learning based on dynamic regularization. ICLR (2021)
2. Campello, V.M., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE TMI **40**(12), 3543–3554 (2021)
3. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7354–7362 (2019)
4. Chapelle, O., Weston, J., Bottou, L., Vapnik, V.: Vicinal risk minimization. NeurIPS **13** (2000)
5. Chen, Z., Zhu, M., Yang, C., Yuan, Y.: Personalized retrogress-resilient framework for real-world medical federated learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 347–356. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_33
6. Duan, M., Liu, D., Chen, X., Liu, R., Tan, Y., Liang, L.: Self-balancing federated learning with global imbalanced data in mobile systems. IEEE Trans. Parallel Distrib. Syst. **32**(1), 59–71 (2020)
7. Eaton-Rosen, Z., Bragman, F., Ourselin, S., Cardoso, M.J.: Improving data augmentation for medical image segmentation. MIDL (2018)
8. Gudur, G.K., Perepu, S.K.: Resource-constrained federated learning with heterogeneous labels and models for human activity recognition. In: Li, X., Wu, M., Chen, Z., Zhang, L. (eds.) DL-HAR 2021. CCIS, vol. 1370, pp. 57–69. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-0575-8_5
9. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: stochastic controlled averaging for federated learning. In: International Conference on Machine Learning, pp. 5132–5143. PMLR (2020)
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv (2013)
11. Lalande, A., et al.: Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI. Data **5**(4), 89 (2020)
12. Li, B., Wu, F., Lim, S.N., Belongie, S., Weinberger, K.Q.: On feature normalization and data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12383–12392 (2021)
13. Li, D., Wang, J.: Fedmd: heterogenous federated learning via model distillation. arXiv preprint arXiv:1910.03581 (2019)
14. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: federated learning on non-iid features via local batch normalization. ICLR (2021)
15. Lin, T., Kong, L., Stich, S.U., Jaggi, M.: Ensemble distillation for robust model fusion in federated learning. NeurIPS **33**, 2351–2363 (2020)
16. Liu, X., Xing, F., El Fakhri, G., Woo, J.: Memory consistent unsupervised off-the-shelf model adaptation for source-relaxed medical image segmentation. Med. Image Anal. 102641 (2022)
17. Liu, X., Yoo, C., Xing, F., Oh, H., El Fakhri, G., Kang, J.W., Woo, J., et al.: Deep unsupervised domain adaptation: a review of recent advances and perspectives. APSIPA Trans. Signal Inf. Process. **11**(1) (2022)
18. Liu, X., et al.: Data augmentation via latent space interpolation for image classification. In: ICPR, pp. 728–733. IEEE (2018)
19. Mancini, M., Porzi, L., Bulo, S.R., Caputo, B., Ricci, E.: Boosting domain adaptation by discovering latent domains. In: CVPR, pp. 3771–3780 (2018)

20. Maria Carlucci, F., Porzi, L., Caputo, B., Ricci, E., Rota Bulo, S.: Autodial: automatic domain alignment layers. In: ICCV, pp. 5067–5075 (2017)
21. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-Efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282. PMLR (2017)
22. Pati, S., Bakas, S.: The federated tumor segmentation (fets) challenge (2021)
23. Qi, X., Yang, G., He, Y., Liu, W., Islam, A., Li, S.: Contrastive re-localization and history distillation in federated CMR segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention–MICCAI 2022. MICCAI 2022. LNCS, vol. 13435, pp. 256–265. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_25
24. Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., Lu, Z.: Generalized federated learning via sharpness aware minimization. arXiv preprint arXiv:2206.02618 (2022)
25. Rieke, N., et al.: The future of digital health with federated learning. NPJ Digit. Med. **3**(1), 1–7 (2020)
26. Shin, M., Hwang, C., Kim, J., Park, J., Bennis, M., Kim, S.L.: Xor mixup: privacy-preserving data augmentation for one-shot federated learning. arXiv preprint arXiv:2006.05148 (2020)
27. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)
28. Tuor, T., Wang, S., Ko, B.J., Liu, C., Leung, K.K.: Overcoming noisy and irrelevant data in federated learning. In: ICPR, pp. 5020–5027. IEEE (2021)
29. Upchurch, P., et al.: Deep feature interpolation for image content changes. In: CVPR (2017)
30. Wang, X., Jin, Y., Long, M., Wang, J., Jordan, M.: Transferable normalization: towards improving transferability of deep neural networks. arXiv:2019 (2019)
31. Wang, Y., Pan, X., Song, S., Zhang, H., Huang, G., Wu, C.: Implicit semantic data augmentation for deep networks. NeurIPS **32** (2019)
32. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.: Unsupervised data augmentation for consistency training. NeurIPS **33**, 6256–6268 (2020)
33. Yin, Y., et al.: Efficient federated tumor segmentation via normalized tensor aggregation and client pruning. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2021. LNCS, vol. 12963, pp. 433–443. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-09002-8_38
34. Yoon, T., Shin, S., Hwang, S.J., Yang, E.: Fedmix: approximation of mixup under mean augmented federated learning. arXiv preprint arXiv:2107.00233 (2021)
35. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: ICCV, pp. 6023–6032 (2019)
36. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. ICML (2018)
37. Zhang, L., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE TMI **39**(7), 2531–2540 (2020)