Anatomical 3D Style Transfer Enabling Efficient Federated Learning with Extremely Low Communication Costs

Yuto Shibata[1*] **Yasunori Kudo**[1] **Yohei Sugawara**[1]
[1]Preferred Networks
yuto19990715@gmail.com
{ykudo,suga}@preferred.jp

Abstract

In this study, we propose a novel federated learning (FL) approach that utilizes 3D style transfer for the multi-organ segmentation task. The multi-organ dataset, obtained by integrating multiple datasets, has high scalability and can improve generalization performance as the data volume increases. However, the heterogeneity of data owing to different clients with diverse imaging conditions and target organs can lead to severe overfitting of local models. To align models that overfit to different local datasets, existing methods require frequent communication with the central server, resulting in higher communication costs and risk of privacy leakage. To achieve an efficient and safe FL, we propose an Anatomical 3D Frequency Domain Generalization (A3DFDG) method for FL. A3DFDG utilizes structural information of human organs and clusters the 3D styles based on the location of organs. By mixing styles based on these clusters, it preserves the anatomical information and leads models to learn intra-organ diversity, while aligning the optimization of each local model. Experiments indicate that our method can maintain its accuracy even in cases where the communication cost is highly limited ($= 1.25\%$ of the original cost) while achieving a significant difference compared to baselines, with a higher global dice similarity coefficient score of 4.3%. Despite its simplicity and minimal computational overhead, these results demonstrate that our method has high practicality in real-world scenarios where low communication costs and a simple pipeline are required. The code used in this project will be publicly available.

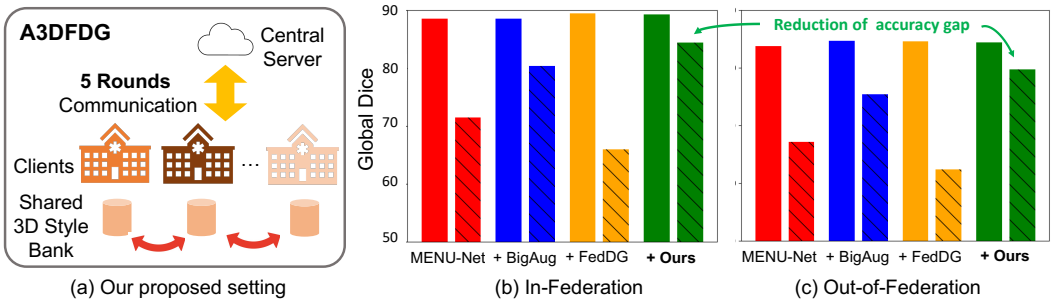
1 Introduction

Recently, the effectiveness of deep learning (DL) in the medical field has been demonstrated through classification and segmentation tasks (16; 9). A large amount of labeled data is required for accurate medical segmentation via DL. This requirement is critical in the field of healthcare because annotating medical images requires high levels of expertise, and collecting pixel-level labels for segmentation tasks is time-consuming and expensive. Furthermore, medical data are highly personal and confidential, making it challenging to share raw data across institutions and countries.

In response to the aforementioned challenges, many studies have aimed to simultaneously protect patient privacy and increase the amount of data available for training by utilizing federated learning (FL) (14; 24; 25). Under the FL scheme, distributed clients perform training using local data and upload their weights to a central server. The central server then aggregates these weights to acquire a more generalized model, whereas all local data are stored under the distributed clients (14).

[*]This work includes contributions made during a summer internship at Preferred Networks.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



(a) Our proposed setting (b) In-Federation (c) Out-of-Federation

Figure 1: (a) Our proposed setting with limitted communication rounds and shared style information. (b)(c) Model accuracy evaluation. The hatched bars indicate the accuracy when the number of communications is reduced from 400 to 5.

Another approach for increasing the amount of annotated data involves combining multiple datasets targeting different organs, resulting in a multi-organ dataset (26; 12). By integrating datasets, the amount of training data increases, resulting in higher accuracy than when training with individual datasets (26). However, implementing FL using these integrated datasets presents a highly challenging issue: **since each local model is optimized in different directions, more iterations are required for the global model to converge, leading to increased overall FL training time and communication cost.** In addition to the well-known domain shift caused by image appearance variation owing to different imaging equipment and protocols (33), differences occur in the targeted organs and the imaging ranges because each dataset was prepared for different purposes. These two sources of domain shift force local models to overfit to their local datasets, significantly slowing down the overall FL training convergence.

Fig 1 illustrates the inefficiency of the existing FL models by demonstrating a significant drop in accuracy with low communication costs (i.e., the number of uploads of the local models to the central server from each local client) while maintaining the total iterations. An in-federation setting means that test data were provided by clients included in the training data, and an out-of-federation setting means that we evaluate the accuracy against unseen clients. The accuracy under the conventional setting with 400 rounds of communications is represented by the plain bars, while the accuracy with five rounds of communications is depicted by the hatched bars. Owing to the aforementioned variation in the optimization directions of local models caused by domain shift, existing models fail to converge and experience a significant drop in accuracy when there is a strict limitation on the number of communications with the central server. This leads to higher operational and communication costs, increases the risk of data leakage when sharing the model, and hinders the system's practical application.

To achieve FL with sufficient scalability and practicality, we propose a novel problem setting, **federated domain generalization with few-round communications**. Toward this goal, we propose **A3DFDG**, **A**natomical **3D** **F**requency **D**omain **G**eneralization for FL (Fig. 1 (a), Fig. 2). This novel method utilizes domain generalization in the frequency space to eliminate differences between domains (clients) and resolve optimization interference among local models. Specifically, we develop a module that successfully extends the existing style adaptations defined in the 2D frequency domain (FDA) (29) to 3D without sharing raw images across clients while utilizing anatomical structural information.

Existing data augmentation methods mix multiple samples randomly, whether the mixing is done in the spatial domain (31; 30) or in the frequency domain (29; 13). However, when dealing with multi-organ datasets, mixing styles obtained from two different organs can lead to the loss of class information contained in each style, potentially distorting the decision boundaries of the model. Therefore, we cluster the styles in the frequency domain using an off-the-shelf organ localization model. Based on these clustered 3D styles, we perform data augmentation in 3D frequency domain, while preserving anatomical information and aligning the optimization of each client's model.

To demonstrate the effectiveness of the proposed method, we conducted evaluations using six datasets under two federated learning (FL) settings: an in-federation setting and an out-of-federation setting. Our results show that even when the number of communications with the central server was significantly reduced to 1.25% of the original setting, we maintained high accuracy comparable to

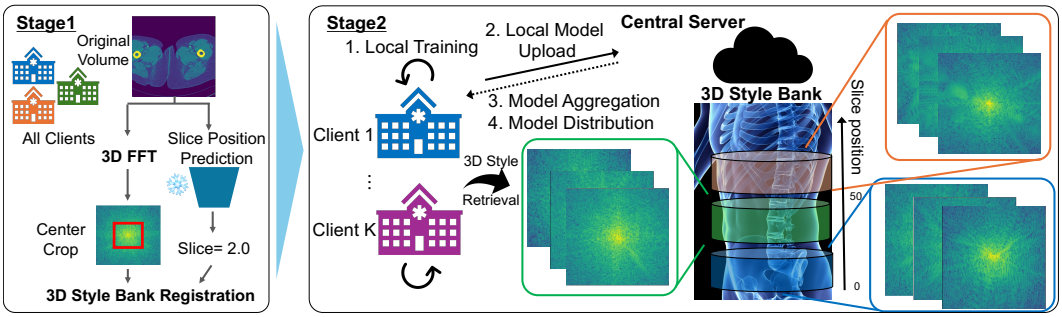


Figure 2: Overview of our A3DFDG. (Stage1) First, we calculate the 3D visual style of each client in the frequency space and store them in clusters based on the predicted slice scores (slice position). (Stage2) During training, we retrieve 3D styles from the same cluster as the samples in the minibatch and perform style transfer without losing organ information.

that of frequent communication settings (Fig 1). Conversely, existing baseline methods are unable to learn multiple organs in a balanced manner and fail to achieve convergence of the global model with limited communication. These findings demonstrate that our model is highly practical for conducting large-scale learning with fewer communications and a simpler overall pipeline.

Table 1 summarizes the communication cost and accuracy when using the recently proposed MENU-Net baseline (size: 6.3GB). Here, our method distributes the 3D style (0.23MB) among clients only once before FL training and these styles are stored at local clients without the need for redistribution during training. Also, as mentioned later, domain generalization is performed using only a part of the frequency spectrums, resulting in much smaller additional communication overhead.

Table 1: Comparison of the trade-off between communication cost and prediction accuracy

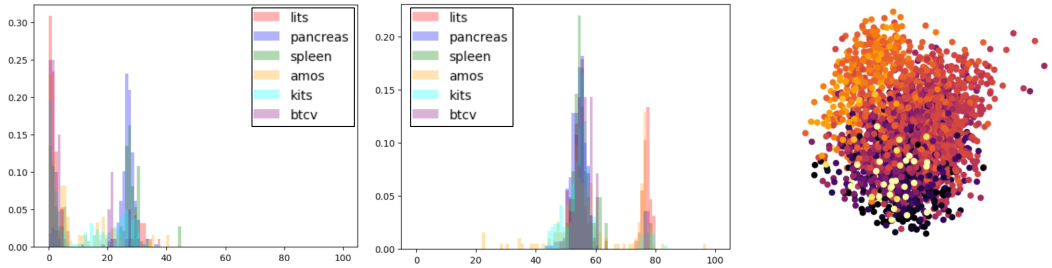
Method	Shared Data × # Rounds	Data traffic	DSC(%)
Existing Work (26)	Model×400	2.5T	88.49
+ Communication Reduction	Model×5	**31.5G**	71.42
+ A3DFDG	Model×5+3D Style×1	**31.5G**	84.38

Our contributions are summarized as follows. (1) We propose the task of medical FL with low communication cost and diverse datasets; (2) we introduce A3DFDG, a domain generalization method utilizing organ structural and low frequency information; (3) we conducted extensive experiments under various FL and communication settings, demonstrating that our simple, yet effective method outperforms existing baselines in settings with limited communication cost.

2 Background and Related Work

Federated Learning. In the FL framework (14), which aims to protect the privacy of patients, distributed learning is conducted without sharing the local data among clients. Clients perform a fixed number of learning iterations using their local datasets, after which the weights of their models are aggregated on a central server. This process involves calculating the weighted average of the local models to obtain a global model (14). The global model is then redistributed to each client for further local training, and this process is repeated until the global model converges.

Domain generalization. In the field of medical imaging, domain shift is a common issue owing to differences in imaging parameters and subject cohorts among hospitals. To address this problem, numerous domain adaptation and generalization methods (both supervised and unsupervised) have been proposed. However, the raw data cannot be shared in the FL framework, strictly prohibiting conventional domain adaptation/generalization approaches. For example, we cannot adopt an instance weighting strategy (22; 3) that requires the similarity scores between the source and target domains because it uses the latent features of each domain. Recently, some studies have addressed domain generalization using the FL scheme. For instance, (8) calculated prototypes that represent each domain



(a) Minimum slice scores (b) Maximum slice scores (c) PCA results for each style

Figure 3: (a)(b) Distribution of the predicted slice scores (= predicted slice position at which images are captured) across six datasets. (c) The distribution of extracted style. Here, color indicates its slice score and PCA is implemented for visualization.

in the feature space, and (23) created synthesized datasets using a generative model to train domain classifiers to obtain domain-invariant features when training local models. However, recent studies have revealed that the latent features of trained models risk privacy data leakage (20) even under collaborative training schemes (5; 28), making these approaches unsuitable in the medical field where strong privacy protections is required. Other studies calculated styles based on FDA (29) to obtain the domain knowledge of each client because the amplitude information of lower frequency bands cannot be used for original image reconstruction without higher frequency and phase components (17; 13). However, to the best of our knowledge, no studies have ever implemented these frequency-based approaches toward 3D medical segmentation tasks with diverse multi-organ datasets.

Multi-organ datasets. Many existing studies that perform learning by combining multiple medical datasets focus on addressing the issue of partial labels in the integrated dataset (18; 26). (18) proposed the exclusion and marginal loss to calculate additional supervision with partial label and (26) tackled the partial label problem under the FL scheme by separating the encoder into sub-encoders to prevent expert models, which undergo supervision with labels, from losing their knowledge by averaging their weights with other non-expert models. Note that this paper addresses the domain shift that arises from merging multiple datasets; thus, handling partial labels falls outside the scope of our proposed method.

Low communication FL. FL requires many communication rounds between a central server and its clients to achieve high accuracy, increasing computational costs and the risk of privacy leakage (32; 4; 15; 21). Furthermore, frequent communication complicates the FL pipeline, making it challenging to handle confidential local data and tmodels. Existing studies have reduced communication costs by decreasing the size of trainable models or sharing the residuals of updates (4; 21). Regarding the reduction in communication frequency, a recent study addressed few-round FL (32; 15). However, DENSE (32) assumes the same task among clients, making it difficult to apply to multi-organ FL. Additionally, these methods have been implemented for relatively simple recognition tasks such as CIFAR-10/100, and they have not been used for 3D medical image segmentation, which requires both high-level and fine-grained visual understanding.

3 Method

Fig. 2 shows the overview of our method. Our proposed framework consists of three parts designed to share the domain information of local datasets while avoiding raw feature leakage and losing anatomical information. Secs. 3.1 and 3.2 describe the style calculation and the style clustering based on organ positions. Sec. 3.3 then explains the training of our models based on these registered style banks. Algorithm 1 presents the detailed steps of our proposed method.

Preliminaries. In this study, K is the number of clients with local datasets. They are expressed as $D = \{D_1, D_2, D_3, \ldots, D_K\}$. Each client contains its local confidential data $D_k = \left\{(x_i^k, y_i^k)\right\}_{i=1}^{N^k}$ while N_k, x_i^k, and y_i^k indicate the size of local data, i-th CT volume, and labels in the k-th dataset, respectively.

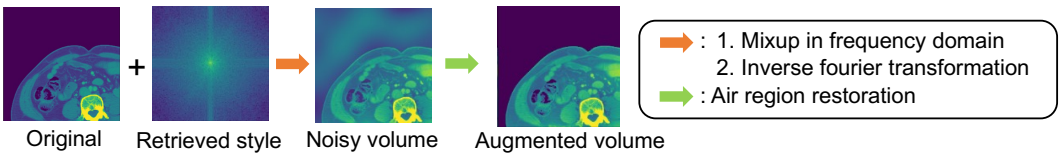


Figure 4: Data augmentation in the frequency domain and subsequent post-processing

3.1 3D style calculation

Previous studies (13; 17; 29) have proved that transferring distribution information in 2D frequency space across clients is effective for domain generalization in the FL training. Toward 3D medical segmentation, our proposed method uses the outputs of a 3D Fourier transform applied to a volume as a style representing each client's domain. First, we respace (up/down-sample) each voxel because the difference in imaging space results in frequency resolution discrepancy when implementing Fourier transformation. We then extract volumes of the same size as used during training and prediction from each voxel at a consistent height interval, thereby creating a 3D style bank keyed by height (the motivation for registration by height is described in Sec. 3.2). The amplitude and phase components of the Fourier transform are denoted as $\mathcal{F}^{\mathcal{A}}$ and $\mathcal{F}^{\mathcal{P}}$, respectively. The style of each local volume x_i^k is expressed as follows:

$$s_i^k(u,v,t) = \mathcal{F}^{\mathcal{A}}(x_i^k)(u,v,t) = |\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sum_{d=0}^{D-1} x_i^k(h,w,d)e^{-j2\pi\left(\frac{hu}{H}+\frac{wv}{W}+\frac{dt}{D}\right)}| \tag{1}$$

where H, W, and D indicate the height, width, and depth of cropped volume, respectively. To preserve the privacy of the patients, we center-crop each Fourier 3D representation, and only the amplitude information in the low-frequency band is registered in the style bank, thus making it impossible to reconstruct the original volume. Additionally, these styles are extracted only from the training data to prevent potential leakage from test data. Following the previous work (13; 29), we calculate weighted sums between different frequency styles to represent continuous and diverse 3D domain information (see Sec. 3.3 for detail).

3.2 Anatomical position registration

Our dataset is a multi-organ dataset created by merging data from multiple clients, each featuring different organs. As a result, while the dataset is large in scale, it covers a wide range of anatomical locations. To investigate the statistics regarding this aspect, we used a pretrained off-the-shelf body part regressor (27). This model outputs slice scores, relative height of the imaging location, where the pelvis and head are set to 0 and 100, respectively. Fig. 3(a) and (b) show the distribution of the maximum and minimum slice scores of the ranges spanned by each volume in the datasets. We can see that different datasets feature different organs, resulting in variations in the imaging locations and potential domain shifts among clients.

To examine the relationship between organ position and domain information defined in the frequency space, we performed dimensionality reduction on the 3D Frequency styles described in Sec. 3.1 using PCA and visualized the results (Fig. 3 (c)). Here, different colors correspond to different estimated slice scores. This visualization demonstrates that styles with the same color tend to cluster together, indicating that our 3D styles encapsulate information related to the position of the organs.

However, with these frequency styles that contain organ position information, randomly mixing the two styles for data augmentation as has been done previously in the spatial or frequency domains (31; 30; 13) can result in the loss of crucial organ information while distorting the decision boundaries of the model. In other words, by mixing these styles during training, the model is optimized to make predictions without relying on the slice position information of the organs, even though this information is actually beneficial for organ identification. Therefore, in this study, we first predict organ locations using a pretrained organ locator model (27) and cluster each style based on predetermined binning thresholds (see 3D Style Bank in Fig. 2). For each cluster, we then perform data augmentation in the frequency domain. This approach forces models to learn intra-organ diversity without imparting biases related to incorrect low-frequency information to the model. Nnote that each

client calculates the slice scores for their data only once before the FL training, and the body position estimator (27) is lightweight, resulting in a minimal additional computation cost for the client.

Algorithm 1 3D Style Bank Registration

Input: Off-the-shelf body part regressor $\phi(w)$, clients $k \in \mathcal{K}$ with local dataset D_k, pre-defined slice score bin size z_{bin} and volume size width w, height h, and depth d

Output: Unified 3D Style Bank B

1: **for** client $k \in \mathcal{K}$ **do**
2: **for** volume $v \in \mathcal{D}_k$ **do**
3: Predict the maximum and minimum slice scores with $\phi(w)$ and calculate volume height $z_{length} = z_{max} - z_{min}$.
4: Random crop sub-volume v of size w, h and d.
5: Calculate the corresponding slice score z' based on z_{min}, z_{length}, and stride size (z-axis) used for cropping.
6: Calculate 3D style s based on Eq. 1 and register it for the style bank B.

$$B[k][z'//z_{bin}].\text{append}(\text{CenterCrop(s)}) \tag{2}$$

7: **end for**
8: **end for**
9: Distribute B across all clients and start FL training utilizing fourier domain generalization (Eq. 4).

3.3 FL training with 3D style bank

Suppose we are training the k-th local model $f_k^t(x;\theta_k^t)$ in round t with the local i-th data x_i^k, where θ_k^t denotes the k-th local model. In each iteration, the local model randomly selects another client k' and retrieves a target 3D style s_{target} that has a similar slice score with the cropped local volume from the precomputed style bank $B[k'][z_{x_i^k}//z_{bin}]$ randomly. During training, the slice score of each cropped sample for style retrieval is calculated based on the maximum/minimum slice score of the original volume and the size of the stride in the z-direction similar to the style bank registration (see Algorithm 1). Subsequently, the two styles are mixed in the frequency space as follows:

$$s' = \alpha \mathcal{F}^{\mathcal{A}}(x_i^k) + (1-\alpha) s_{target} \tag{3}$$

$$x_i^{k'} = \mathcal{F}^{-1}\left([(M_\beta \circ s') + \left((1-M_\beta) \circ \mathcal{F}^{\mathcal{A}}(x_i^k)\right), \mathcal{F}^{P}(x_i^k)]\right) \tag{4}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier Transform, α is the hyperparameter for this MixUp operation, and M is a mask whose value is one at the predefined center region; otherwise, it is zero.

$$M_\beta(h,w,d) = \mathbf{1}_{(h,w,d) \in [\![-\beta_h H:\beta_h H],[-\beta_w W:\beta_w W],[-\beta_d D:\beta_d D]\!]} \tag{5}$$

where β is a hyperparameter that controls the extent of the style transfer. Note that since β is very small (Sec. 4), the communication cost for sharing these styles is limited (Table 1).

Inverse Fourier Transformation after style mixing results in artifacts in the external regions of the body as shown in Fig 4. Therefore, we preserve anatomical information, such as the distance to the contour of the body using a threshold-based air mask for post-refinement. We use air threshold $\tau_{air} = -200$ to filter air pixels and those pixels are filled with the original value after style transformation.

Each time local training is completed, we obtain the global model by calculating the weighted average of each local model based on their dataset sizes following common FL implementation (14; 26).

$$\theta^{t+1} = \sum_k \frac{|D_k|}{\sum_j |D_j|} \cdot \theta_k^t. \tag{6}$$

4 Experimental Settings

Datasets and Preprocessing. Following previous work (26), we used six (multi-) organ segmentation datasets: 1) the liver tumor segmentation challenge (2) (LiTs), 2) kidney tumor segmentation challenge (KiTS) (7; 6), 3) pancreas, 4) spleen segmentation datasets in medical segmentation decathlon challenge(19; 1), 5) multi-modal abdominal multi-organ segmentation challenge (AMOS)(10), and 6)

Table 2: Dice similarity coefficient (DSC) scores in the in-federation setting.

		DSC (%)					
Model	# Rounds	Liver	Kidney	Pancreas	Spleen	Gallbladder	Global
FedAvg (14)	400	**94.95**	94.62	81.40	92.13	80.24	88.67
	5	91.29	90.28	58.75	75.07	43.46	71.77
MENU-Net (26)	400	94.43	94.85	81.97	92.60	78.6	88.49
	5	92.29	91.64	76.44	28.92	67.81	71.42
MENU-Net	400	93.94	94.25	**82.08**	91.70	80.62	88.52
+ BigAug (26; 34)	5	91.19	90.00	75.12	80.63	64.43	80.28
MENU-Net	400	94.39	**94.87**	81.77	**93.18**	**82.77**	**89.40**
+ FedDG (26; 13)	5	**92.52**	80.81	**77.79**	8.51	69.89	65.90
Ours	400	94.57	94.61	81.64	93.02	82.21	89.21
	5	92.51	**92.59**	77.17	**89.23**	**70.42**	**84.38**

multi-atlas labeling beyond the cranial vault challenge (BTCV) (11) datasets. These datasets contain 131, 210, 281, 41, 200, and 30 volumes, respectively. For more detailed information about these datasets please refer to the prior work (26). We also implemented the same preprocessing as in (26) for downsampling and pixel normalization with clipping.

Baselines. We compared our proposed model with the following baseline models: (i) FedAvg (14), the original work on FL that calculates the average of weights after local training; (ii) MENU-Net (26), which separates the encoder into multiple sub-encoders to prevent model optimization interference during the global model update; (iii) BigAug (33), which utilizes a set of heavy augmentations to generalize the model towards unknown domains; (iv) FedDG (13), the closest research to our work, and it achieves highly generalized Federated Learning while preserving privacy by center-cropped frequency spectrums. The 3D organ datasets we handle are difficult to collect, and some clients have limited local data (e.g., the spleen segmentation dataset (1) contains only 24 training samples). Therefore, instead of adopting the meta-learning approach proposed in the FedDG paper, we adopted only the data augmentation part toward the multi-source domain based on continuous frequency space interpolation. Regarding (iii) (iv), MENU-Net was adopted as the network architecture. In addition, (i) uses the same 3D convolutional layers as MENU-Net except for sub-encoders. For more detailed hyperparameter settings, please refer to the Supplementary Material.

Evaluation Metrics. We calculated the average dice similarity coefficient score (DSC) and average symmetric surface distance (ASD) for each organ. The macro average across clients was calculated in an in-federation setting. Also, the global accuracy was calculated using the macro average across all organs.

Implementation Details. We adopted the MENU-Net architecture for the model of our proposed method and trained our model with dice, cross-entropy, and marginal and exclusion loss functions following (26). Regarding the hyperparameters, α is randomly sampled within $[0.0, 1.0]$, and we set $\beta_w, \beta_h,$, and β_d to 0.01, 0.01, and 0.05, respectively. These hyperparameters are used for both of our method and FedDG (13). We used 10% and 30% of each local data for validation and testing while all images in BTCV were used for out-of-federation testing. The training and testing batch sizes were set to four and two, respectively. The learning rate was set to 0.01 for 400 communication rounds setting and 0.001 for 5 rounds settings to stabilize the training processes.

5 Experimental Results.

To evaluate the efficacy of our model and the other baselines, we trained and evaluated them with two communication frequencies under two distinct settings: (i) an in-federation setting and (ii) an out-of-federation setting.

Table 3: DSC scores in the out-of-federation setting.

Model	# Rounds	DSC (%)					
		Liver	Kidney	Pancreas	Spleen	Gallbladder	Global
FedAvg (14)	400	**95.03**	**88.57**	79.10	90.99	67.84	84.31
	5	92.20	87.89	53.07	66.59	32.42	66.43
MENU-Net (26)	400	94.83	87.87	77.85	89.19	**68.73**	83.69
	5	94.41	87.65	74.45	33.23	45.78	67.10
MENU-Net +BigAug (26; 34)	400	94.57	87.77	79.58	90.72	70.70	**84.67**
	5	93.98	85.44	72.20	81.04	44.13	75.36
MENU-Net + FedDG (26; 13)	400	94.95	88.12	78.77	**92.27**	68.59	84.54
	5	94.15	78.40	75.92	6.84	**56.59**	62.38
Ours	400	94.76	87.68	**79.65**	91.67	68.21	84.39
	5	**94.64**	**89.35**	**75.98**	**88.79**	49.56	**79.66**

Table 4: ASD scores in the out-of-federation setting.

Model	# Rounds	ASD (mm)					
		Liver	Kidney	Pancreas	Spleen	Gallbladder	Global
FedAvg (14)	400	2.57	**4.27**	**1.98**	1.67	2.60	2.62
	5	3.91	6.08	38.15	5.18	11.18	12.90
MENU-Net (26)	400	**2.53**	4.74	2.66	2.43	2.21	2.91
	5	3.33	6.41	4.70	166.84	**2.55**	36.77
MENU-Net + BigAug (26; 34)	400	2.83	4.95	2.23	2.22	2.00	2.84
	5	2.74	7.64	12.25	**3.61**	5.85	6.42
MENU-Net + FedDG (26; 13)	400	2.60	4.90	2.69	**1.40**	**1.22**	**2.56**
	5	2.66	9.24	3.22	26.16	8.41	9.94
Ours	400	2.71	5.31	2.49	2.05	3.99	3.31
	5	**2.49**	**4.69**	**2.59**	4.34	3.91	**3.60**

5.1 Comparison with Other Baselines

Table 2 presents a quantitative comparison in the in-federation setting. We can see that the methods using domain generalization in frequency space recorded high accuracy under the frequent communication setting (Rounds= 400). However, only the proposed method is able to maintain high accuracy even when the number of communications with the central server is significantly reduced to 1.25 % of the original cost (Rounds= 5). Also, it can be observed that many previous studies face challenges in maintaining accuracy for all organs when the number of communications is restricted. For example, while FedDG (13) achieves high accuracy for all organs with frequent communications, in the limited communication setting, the training for the spleen and kidney has not converged, resulting in significantly lower accuracy for these organs.

Table 3 presents the quantitative results in the out-of-federation setting. Similar to the in-federation setting, accuracy reduction is limited ($< 5\%$) even when the number of rounds is significantly reduced while other baseline methods significantly degraded their accuracy (-18%, -17%, -8%, -22%), indicating that our domain generalization method enables efficient FL training while reducing the optimization interference among local models.

Table 4 shows the ASD scores in the out-of-federation setting. The accuracy of our proposed method is significantly higher than that of existing methods in realistic settings under low communication costs while other models have very unstable training processes and fail to converge. For the results in the in-federation setting, please refer to our supplementary material.

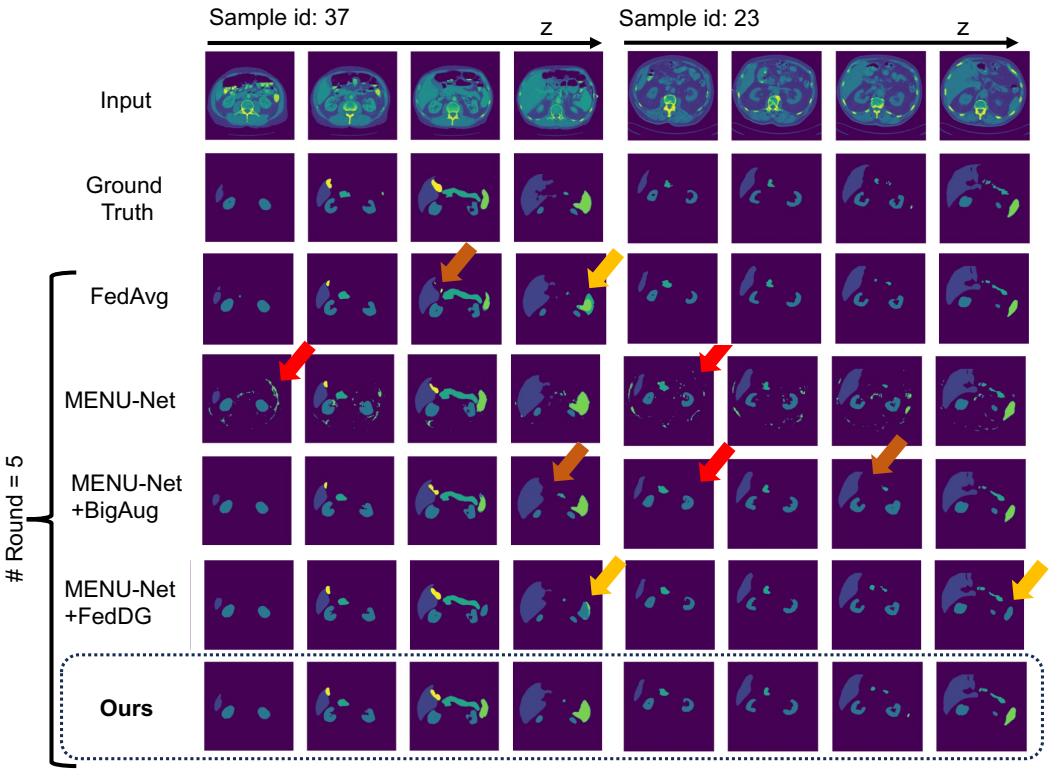


Figure 5: Qualitative results in the out-of-federation setting.

These results suggest the following. First, when communication rounds are limited, performing heavy data augmentation (ours, BigAug (33)) achieves efficient model aggregation. Alternatively, when random mixup is applied as in FedDG (13), incorrect biases can be introduced into the local models, leading to lower accuracy for some organs compared to when no data augmentation is applied.

5.2 Qualitative results

Fig 5 shows the qualitative results in an out-of-federation setting. These results included two patients, and we displayed the results by slicing at equal intervals in the z-direction. The existing methods (14; 26; 33; 13) produce many false positives (red arrows), false negatives (brown arrows), and misclassification (yellow arrows) in a low communication setting. By contrast, our proposed method maintains high accuracy even when the number of communications is restricted.

5.3 Ablation study

We investigated the effects of our technical contributions, including position-based style clustering (slice score matching) and post-processing with an air mask. Table 5 presents the accuracy in the in-federation and the out-of-federation setting respectively. Based on these results, we can see that both our proposed modules contributed to improving the estimation accuracy compared with the scores in Table 2, specifically in a low communication setting.

5.4 Model-Agnostic Efficacy of our method

Tables 6 demonstrates the global accuracy in both settings when we apply our method for the FedAvg (14) model. We can observe that our method significantly improved accuracy at lower communication costs, demonstrating that it has a minimal dependency on model architecture choice and a high potential to be utilized as a plug-in function for various model architectures.

Table 5: Ablation study

Model	#Rounds	In-Fed DSC (%)	Out-of-Fed DSC (%)
Ours	400	88.69 (-0.52)	83.51 (-0.88)
w/o slice score matching	5	73.01 (-11.37)	65.97 (-13.68)
Ours	400	87.90 (-1.3)	84.69 (+0.30)
w/o contour preservation	5	83.01 (-1.37)	78.16 (-1.5)

Table 6: Model-agnostic efficacy

Model	# Rounds	In-Fed DSC(%)	Out-of-Fed DSC(%)
FedAvg (14)	400	88.67	84.31
	5	71.77	66.43
FedAvg (14)+A3DFDG	400	88.62 (-0.05)	84.07 (-0.28)
	5	73.71 (+1.94)	68.56 (+2.13)

6 Discussion and Limitations

Although our proposed method significantly improves accuracy in settings with limited communication costs, the improvement margin is limited in scenarios where frequent communication is possible. Moreover, one of the current main limitations of our framework is that each client needs to calculate the height of local volumes using a pre-trained organ position estimator beforehand. In future work, we plan to address this by using segmentation predictions to determine the volume occupied by each organ and dynamically calculating the corresponding slice score on the fly.

7 Conclusion

This paper propose A3DFDG, an Anatomical 3D Frequency Domain Generalization method to achieve efficient FL for a heterogeneous multi-organ dataset. Compared with existing methods that randomly sample and mix styles, the proposed method utilizes 3D styles clustered based on the organ location. This approach enables domain generalization without compromising anatomical information and forces models to learn intra-organ diversity. Despite its simplicity and minimal computational overhead, our method maintains accuracy with restricted communication frequency while existing methods significantly decrease in accuracy or fail to converge. We believe that this work offers a new possibility for highly practical large-scale FL with limited communication costs and diverse data.

Acknowledgements. We would like to express our gratitude to the medical imaging team at Preferred Networks for the valuable discussions and helpful feedbacks, and to the cluster team for enabling the execution of numerous experiments.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 6, 7

[2] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 6

[3] Veronika Cheplygina, Isabel Pino Peña, Jesper Holst Pedersen, David A. Lynch, Lauge Sørensen, and Marleen de Bruijne. Transfer learning for multicenter classification of chronic obstructive pulmonary disease. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1486–1496, 2018. 3

[4] Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. Learning federated visual prompt in null space for mri reconstruction. In *CVPR*, pages 8064–8073, 2023. 4

[5] Zecheng He, Tianwei Zhang, and Ruby B Lee. Model inversion attacks against collaborative inference. In *ACSAC*, pages 148–162, 2019. 4

[6] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. 6

[7] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 6

[8] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322. IEEE, 2023. 3

[9] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. 1

[10] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *NeurIPS*, 35:36722–36732, 2022. 6

[11] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015. 7

[12] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21152–21164, October 2023. 2

[13] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 2, 4, 5, 7, 8, 9

[14] 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. PMLR, 2017. 1, 3, 6, 7, 8, 9, 10

[15] Younghyun Park, Dong-Jun Han, Do-Yeon Kim, Jun Seo, and Jaekyun Moon. Few-round learning for federated learning. *NeurIPS*, 34:28612–28622, 2021. 4

[16] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[17] Donald Shenaj, Eros Fanì, Marco Toldo, Debora Caldarola, Antonio Tavera, Umberto Michieli, Marco Ciccone, Pietro Zanuttigh, and Barbara Caputo. Learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 444–454, 2023. 4, 5

[18] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021. 4

[19] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 6

[20] Abhishek Singh, Ayush Chopra, Ethan Garza, Emily Zhang, Praneeth Vepakomma, Vivek Sharma, and Ramesh Raskar. Disco: Dynamic and invariant sensitive channel obfuscation for deep neural networks. In *CVPR*, pages 12125–12135, 2021. 4

[21] Rui Song, Liguo Zhou, Lingjuan Lyu, Andreas Festag, and Alois Knoll. Resfed: Communication efficient federated learning by transmitting deep compressed residuals. *arXiv preprint arXiv:2212.05602*, 2022. 4

[22] Christian Wachinger, Martin Reuter, Alzheimer's Disease Neuroimaging Initiative, et al. Domain adaptation for alzheimer's disease diagnostics. *Neuroimage*, 139:470–479, 2016. 3

[23] Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. Dafkd: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20412–20421, 2023. 4

[24] Yawen Wu, Dewen Zeng, Zhepeng Wang, Yiyu Shi, and Jingtong Hu. Federated contrastive learning for volumetric medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 367–377. Springer, 2021. 1

[25] An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger R Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20866–20875, 2022. 1

[26] Xuanang Xu, Hannah H Deng, Jamie Gateno, and Pingkun Yan. Federated multi-organ segmentation with inconsistent labels. *IEEE Transactions on Medical Imaging*, 2023. 2, 3, 4, 6, 7, 8, 9

[27] Ke Yan, Le Lu, and Ronald M Summers. Unsupervised body part regression via spatially self-ordering convolutional neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1022–1025. IEEE, 2018. 5, 6

[28] Mengda Yang, Ziang Li, Juan Wang, Hongxin Hu, Ao Ren, Xiaoyang Xu, and Wenzhe Yi. Measuring data reconstruction defenses in collaborative inference systems. *NeurIPS*, 35:12855–12867, 2022. 4

[29] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020. 2, 4, 5

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2, 5

[31] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 5

[32] Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. Dense: Data-free one-shot federated learning. *NeurIPS*, 35:21414–21428, 2022. 4

[33] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020. 2, 7, 9

[34] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020. 7, 8