One Model to Unite Them All: Personalized Federated Learning of Multi-Contrast MRI Synthesis

Onat Dalmaz, Muhammad U Mirza, Gökberk Elmas, Muzaffer Özney, Salman UH Dar, Emir Ceyani, Salman Avestimehr, Fellow, Tolga Çukur*, Senior Member

Abstract—Multi-institutional collaborations are key for learning generalizable MRI synthesis models that translate source-onto target-contrast images. To facilitate collaboration, federated learning (FL) adopts decentralized training and mitigates privacy concerns by avoiding sharing of imaging data. However, FL-trained synthesis models can be impaired by the inherent heterogeneity in the data distribution, with domain shifts evident when common or variable translation tasks are prescribed across sites. Here, we introduce the first personalized FL method for MRI Synthesis (pFLSynth) to improve reliability against domain shifts. pFLSynth is based on an adversarial model that produces latents specific to individual sites and source-target contrasts, and leverages novel personalization blocks to adaptively tune the statistics and weighting of feature maps across the generator stages given latents. To further promote site specificity, partial model aggregation is employed over downstream layers of the generator while upstream layers are retained locally. As such, pFLSynth enables training of a unified synthesis model that can reliably generalize across multiple sites and translation tasks. Comprehensive experiments on multi-site datasets clearly demonstrate the enhanced performance of pFLSynth against prior federated methods in multi-contrast MRI synthesis.

Index Terms—federated learning, personalization, MRI, synthesis, translation, heterogeneity, domain shift

I. INTRODUCTION

MRI offers non-invasive assessment of anatomy with rich diagnostic information accumulated over multiple tissue contrasts [1]. Yet, it suffers from prolonged scan times due to its limited signal-to-noise ratio (SNR) efficiency. Costs associated with multi-contrast protocols can prohibit comprehensive acquisitions or repeat runs of corrupted contrasts during an exam [2]. Contrast translation is a promising solution wherein missing images in the protocol are synthesized from a subset of acquired images [3]. Adoption of deep learning has paved the way for centrally-trained models that offer leaps in synthesis performance [4]–[7]. Unfortunately, training of generalizable models requires diverse datasets that are difficult to curate centrally due to patient privacy risks [8].

This study was supported in part by a TUBITAK BIDEB scholarship awarded to O. Dalmaz, and by TUBA GEBIP 2015 and BAGEP 2017 fellowships awarded to T. Çukur (Corresponding author: Tolga Çukur). O. Dalmaz, U. Mirza, G. Elmas, M. Ozney, SUH. Dar and T. Çukur are with the Department of Electrical and Electronics Engineering, and the National Magnetic Resonance Research Center, Bilkent University, Ankara, Turkey (e-mail: {onat, mumirza, gokberk, muzaffer, salman, cukur}@ee.bilkent.edu.tr). E. Ceyani and S. Avestimehr are with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: {ceyani@, avestime@})usc.edu).

Federated learning (FL) addresses this limitation based on decentralized model training across multiple institutions [9]–[14]. An FL server sporadically aggregates locally-trained models to compute a shared global model [15], [16], and then broadcasts the global model onto each site for further training. Thus, collaborative training is performed on diverse datasets from multiple sites without sharing of imaging data [17]. However, the resultant models can be impaired by the data heterogeneity naturally evident in multi-site datasets [18], [19]. Previous studies on FL-based medical imaging have introduced prominent approaches to cope with data heterogeneity in segmentation [14], [20]–[24], classification [13], [25], [26], and reconstruction [27]–[29] tasks. However, to our knowledge, no prior study has attempted to address data heterogeneity in federated MRI synthesis.

Learning-based MRI synthesis recovers unavailable target-contrast images of an anatomy provided as input acquired source-contrast images [30], [31]. Models are trained on datasets comprising samples from the desired source-target configuration [32]. In a restricted setup, a multi-site model can be built for a single task with a common configuration (e.g., $T_1\rightarrow T_2$ at all sites where source$\rightarrow$target), resulting in implicit heterogeneity across sites due to variations in sequence parameters or imaging hardware [27], [29]. In a more flexible setup, a multi-site model can be built for multiple tasks with variable configuration across sites (e.g., $T_1\rightarrow T_2$ in some, $T_2\rightarrow$PD in others). This will result in explicit heterogeneity due to differences in the identity of source and target contrasts. Both implicit and explicit heterogeneity can induce notable performance losses in synthesis models [7].

Here, we introduce a novel personalized FL method for MRI Synthesis (pFLSynth) that effectively addresses implicit and explicit heterogeneity in multi-site datasets. pFLSynth employs a unified adversarial model that produces latent variables specific to individual sites and source-target contrasts. To cope with heterogeneity, novel personalization blocks are introduced that adaptively modulate feature maps across the generator given these site- and contrast-specific latents. For further personalization, we propose partial network aggregation on downstream generator layers, while upstream layers are kept local. These design elements enable pFLSynth to reliably generalize across multiple sites and synthesis tasks with a unified model. Comprehensive experiments on multi-site MRI data clearly demonstrate the superior performance of pFLSynth against prior federated models. Code for pFLSynth is available at: https://github.com/icon-lab/pFLSynth.
Contributions

- We introduce the first personalized FL method for MRI synthesis to improve performance and flexibility in multi-site collaborations.
- To perform diverse tasks with a unified model, our personalized architecture adaptively modulates feature maps to each site and source-target configuration.
- We propose partial network aggregation on downstream layers to maintain site generality, while upstream layers are kept local to maintain site specificity.

II. RELATED WORK

Multi-contrast MRI synthesis has witnessed a recent surge in deep learning methods [33]–[35]. Earlier studies proposed convolutional neural networks (CNN) with pixel-wise loss terms [4]–[6]. To improve capture of tissue details, generative adversarial networks (GAN) were also introduced based on adversarial loss terms that indirectly learn the distribution of target images [7], [36]–[40]. Commonly, maximal performance has been aimed by training a singular model for each source-target configuration [41]. To mitigate computational burden, later studies have instead proposed unified models capable of performing multiple tasks [42]–[46]. However, despite demonstrated success in learning-based synthesis, prior studies have predominantly reported centralized models that require cross-site transfer of imaging data [8].

To remedy privacy concerns, FL conducts decentralized training via communicating model parameters [17]. Given their potential for collaborative studies, FL methods have been proposed for several imaging tasks including segmentation [9], [10], [14], [20], classification [12], [21], [25], [26], reconstruction [27]–[29], [47], and unconditional image generation [48]. Yet, the potential of decentralized procedures in multi-modal medical image translation remains largely unexplored. FL has recently been adopted for multi-contrast MRI synthesis based on cycle-consistent models [49]. Promising results were reported for training singular models built independently for each separate source-target configuration. Yet, to our knowledge, no prior FL study has attempted to address heterogeneity in multi-site datasets during MRI synthesis.

Here, we introduce the first FL method for personalized MRI synthesis that addresses heterogeneity existent for both common and variable source-target configurations across sites (Fig. 1). To do this, pFLSynth introduces novel personalization blocks in the generator of an adversarial model (Fig. 2). Receiving site- and contrast-specific latent variables, these blocks adaptively tune the statistics of feature maps via adaptive instance normalization [51], and tune the attributed importance of feature channels via adaptive channel weighting. Several recent studies have considered adaptive normalization for imaging tasks [47], [52]–[55]. In [52], normalization was used for ultrasound imaging to switch among various output types in a centralized model. In [47], [53], [54], normalization was used to lower model complexity in a cycle-consistent architecture. In [55], normalization was used to synthesize images at intermediate echo times (TE) via a centralized model. Unlike prior studies, we employ normalization layers to personalize a federated model to each individual site and source-target configuration, and couple them with novel channel weighting layers for improved specificity.

pFLSynth employs shared generators across sites, albeit the generator is split centrally in its residual bottleneck to enable partial aggregation. A recent study on MRI reconstruction has proposed sharing the encoder in a UNet backbone while keeping the decoder unshared to enhance site specificity in encoded representations [28]. In contrast, here we propose to share the downstream layers of the residual generator while keeping upstream layers local to maintain specificity in encoded representations. Taken together, these unique aspects enable pFLSynth to effectively address heterogeneity in multi-contrast data for reliable MRI synthesis.

III. THEORY

A. MRI Synthesis with Adversarial Models

Adversarial models have become pervasive in MRI synthesis due to their sensitivity for high-frequency features [7], [43], [56], [57]. For adversarial learning, a generator $G$ synthesizes a target image ($\hat{x}_t = G(x_s)$) given as input a source image ($x_s$), while a discriminator $D$ distinguishes actual ($x_t$) and synthetic ($\hat{x}_t$) target images. Assuming spatially-registered images, a GAN is typically trained to minimize:

$$
\mathcal{L}_{syn}(D, \theta) = \mathbb{E}_{x_s, x_t}[-(D(x_t) - 1)^2 - D(G(x_s))^2 + \lambda_{pix}||x_t - G(x_s)||_1],
$$

where $\mathbb{E}$ denotes expectation, $D$ are training data comprising source-target images, $\theta = \{\theta_G,\theta_D\}$ are model parameters, the first two terms reflect an adversarial loss, the last term reflects a pixel-wise loss with relative weight $\lambda_{pix}$. The traditional approach uses centralized training, where data samples from multiple institutions are aggregated in a single repository [8]. This, however, introduces privacy risks for patients.

Alternatively, decentralized training can be performed via communication between an FL server hosting a global generator ($G$ with $\theta_G$), and sites keeping local copies ($G^k$ for site $k \in \{1, ..., K\}$) [49]. Discriminators are typically unshared across sites to minimize risk of information leak [58]. In each communication round, local copies are initialized with the global model transmitted by the server ($\theta^*_G \leftarrow \theta_G$). Local models are then trained to minimize a local synthesis loss:

$$
(\theta^*_G, \theta^*_D) = \operatorname{argmin}_{G^k} \mathcal{L}_{syn}^k(D^k, \theta^k),
$$

where $D^k$ are training data ($x^k_s, x^k_t$), and ($s_c, t_c$) denotes the $c$th source-target configuration at site $k$ ($c \in \{1, ..., C\}$). After each round, local models are aggregated on the server [15]:

$$
\theta_G = \sum_{k=1}^{K} \alpha^k \theta^k_G,
$$

where $\alpha^k$ denote relative site weights typically set to $\frac{n^k}{\sum n^k}$, where $n$ is the total number of training samples and $n^k$ is the number of training samples at site $k$. The trained global model ($G_{G^*}$) is eventually used for local inference:

$$
\hat{x}^*_t = G^*(x^*_s).
$$

Domain shifts due to implicit and explicit data heterogeneity will lead to a compromise model among various sites and synthesis tasks, leading to suboptimal performance.
B. Personalized Federated Learning of MRI Synthesis

For improved performance in multi-site studies, we propose a personalized FL method for MRI synthesis, pFLSynth (Fig. 1). pFLSynth leverages novel personalization blocks (PB) composed of Adaptive Instance Normalization (AdaIN) [51] and novel Adaptive Channel Weighting (AdaCW) layers (Fig. 2). Partial network aggregation is adopted to further improve site specificity and communication efficiency.

**B.1 Personalized Architecture:** pFLSynth is an adversarial model that receives a source image along with site and source-target configuration information as input. The generator employs a mapper that produces latent variables \( w \) for modulating feature maps and a synthesizer for source-to-target translation.

**Mapper (M):** \( M \) is an \( L_M \)-layer multi-layer perceptron (MLP). Receiving a binary vector for site index \( (v^k \in \mathbb{Z}_2, \mathbb{Z}_2 = \{0, 1\}) \) and a binary vector for indices of source and target contrasts \( (u_c \in \mathbb{Z}_2^{2C}) \), it produces a latent vector:

\[
 w^k_c = M(v^k \oplus u_c),
\]

where \( w^k_c \in \mathbb{R}^J \), \( \oplus \) is concatenation, \( J \) is latent dimensionality. Parametrized with \( \theta_M \), \( M \) produces site- and task-specific latents to drive PBs for PBs.

**Synthesizer (S):** Receiving as input a source image \( x_s \) and latent vector \( w^k_c \), \( S \) generates a target image \( \hat{x}_t \). The backbone is inspired by the ResNet model [7], [59] with a residual bottleneck between an encoder and a decoder (see Fig. 2). For each site and source-target configuration, synthesizer feature maps elicit divergent statistics across spatial and channel dimensions [60]. To mitigate this heterogeneity, we introduce novel PBs inserted after each convolutional block in the synthesizer except for the final decoder block. Let \( \{S_1, S_2, \cdots, S_{L_S}\} \) with parameters \( \theta_{S_1, \cdots, S_{L_S}} \) be the set of synthesizer stages. At the \( i \)-th stage, input feature maps \( f_{i-1} \in \mathbb{R}^{F_{i-1}, H_{i-1}, W_{i-1}} \) (\( F_{i-1}, H_{i-1}, W_{i-1} \) are the number of channels, height and width) are processed via \( CB_i \):

\[
g_i = CB_i(f_{i-1}) \in \mathbb{R}^{F_i, H_i, W_i},
\]

where \( CB_i \) denotes the \( i \)-th convolutional block.

\( PB_i \) first transforms \( w^k_c \) into scale and bias vectors \( \gamma_i, \beta_i \in \mathbb{R}^{F_i} \), and then normalizes the statistics of \( g_i \):

\[
\gamma_i(k, c) = Q^g_i w^k_c + b^g_i; \beta_i(k, c) = Q^\beta_i w^k_c + b^\beta_i,
\]

where \( Q^g_i, Q^\beta_i \in \mathbb{R}^{F_i \times J} \) and \( b^g_i, b^\beta_i \in \mathbb{R}^{F_i} \) are learnable projections, \( J \in \mathbb{R}^{H_i, W_i} \) is a matrix of ones, \( \mu, \sigma \) compute the mean and standard deviation of individual channels \( g_i \in \mathbb{R}^{H_i, W_i} \).

Next, a novel AdaCW layer weights feature channels according to a site-task relevance vector \( CW_i \in \mathbb{R}^{F_i} \), produced by an \( L_{CW} \)-layer MLP given the latents \( w^k_c \):

\[
CW_i(k, c) = MLP(w^k_c)
\]

where \( \odot \) denotes Hadamard product, and \( f_i \) is the output feature map. The overall mapping through the synthesizer is a cascade of projections through CBs and PBs:

\[
\hat{x}_t = CB_{L_S} \circ PB_{L_S-1} \circ CB_{L_S-1} \cdots \circ PB_1 \circ CB_1(x_s, w^k_c),
\]

where \( \circ \) denotes functional composition, and \( S_i := PB_i \circ CB_i \) for \( i < L_s \) while \( S_{L_S} := CB_{L_S} \) (i.e., the final decoder block).
Partial Network Aggregation (PNA): S is split at stage $L_U$ into two disjoint subsets $S_U = \{S_1, \ldots, S_{L_U}\}$ and $S_D = \{S_{L_U+1}, \ldots, S_{L_S}\}$ where $L_U$ is selected from $\{1, \ldots, L_S\}$. The set of upstream stages $S_U$ with parameters $\theta^k_{S_U}, k \in \{1, \ldots, K\}$ are kept locally at each site. In contrast, the set of downstream stages $S_D$ with parameters $\theta^k_{S_D}$ are shared. Thus, pFLSynth only aggregates $S_D$ to improve site specificity in source-image representations. Here, PNA is only exercised on CBs within S. While it can also be adopted for PBs, we observed that using local PBs persistently throughout S performs similarly. Thus, to lower communication costs and potential for information leakage, all PBs are kept locally.

Discriminator ($D$): A local discriminator $D^k$ with parameters $\theta^k_D$ is trained at site $k$ based on a conditional patch-based architecture [7]. Given the source image, $D^k$ estimates the probability that $x$ is an actual target image:

$$p_D = D(x, x_s),$$

where $x$ can be an actual or synthetic target image.

B.2 Federated Training: Decentralized learning is performed for $P$ communication rounds between the server and individual sites (Alg. 1). The downstream synthesizer $S_D$ and mapper $M$ are shared across sites, albeit the upstream synthesizer $S_U$ and discriminator $D$ are unshared. In the first round, the server randomly initializes a global generator $G = \{S_D, M\}$ with parameters $\theta_G = \{\theta^0_{S_D}, \theta^0_M\}$. At the start of each round, the server broadcasts the global $S_D$ and $M$ to the sites:

$$\theta^0_{S_D} \leftarrow \theta^0_{S_D}; \quad \theta^0_M \leftarrow \theta^0_M; \quad k = 1, 2, \ldots, K.$$

Local $S_U$ and $D$ are set to their states in the previous round:

$$\theta^k_{S_U} \leftarrow \theta^k_{S_U}; \quad \theta^k_D \leftarrow \theta^k_D,$$

A local generator is then formed as:

$$G^k = \{(S^k_U \cup S^k_D), M^k\}.$$

Next, each local generator is trained for $E$ epochs. Note that pFLSynth consolidates different synthesis tasks within and across sites. Thus, local training data $D^k$ comprise multiple source-target configurations at site $k$:

$$\text{config}^k = \{\{s_1, t_1\}, \ldots, \{S_{c}, t_c\}\},$$

where a fixed number of configurations $C$ is assumed at each site. Given $\hat{x}^k_{tc} = G^k(x^k_{ts}, v^k_{uc})$, local models are trained to minimize a compound local synthesis loss across tasks:

$$L^k_{syn}(D^k, \theta^k) = \sum_{c=1}^{C} \mathbb{E}_{x^k_{ts}, x^k_{ts}} \left[-(D^k(x^k_{tc}, x^k_{ts}) - 1)^2 - D^k(x^k_{tc}, x^k_{ts})^2 + \lambda_{pix} ||x^k_{ts} - x^k_{tc}||_1\right].$$

At the end of a round, each site sends its downstream synthesizer and mapper to the server for aggregation [61]:

$$\theta^k_{S_D} = \frac{1}{K} \sum_{k=1}^{K} n^k_{S_D}, \quad \theta^k_M = \frac{1}{K} \sum_{k=1}^{K} n^k_{M}.$$  

During inference, each site forms a personalized generator as in Eq. 15 to perform contrast translation:

$$\hat{x}^k_{tc} = G^k(x^k_{ts}, v^k_{uc}).$$

Note that our federated generator performs an adaptive source-to-target mapping at each site and for each synthesis task.

IV. METHODS

A. Datasets

Experiments were conducted on four multi-contrast MRI datasets: IXI \footnote{https://brain-development.org/ixi-dataset/} , BRATS [62], MIDAS [63], and OASIS [64]. IXI and MIDAS contain data from healthy subjects, BRATS contains data from glioma patients, and OASIS contains data from subjects with cognitive decline. Each dataset was treated as a separate site in the FL setup. Subjects within each site were split into non-overlapping training, validation, and test sets. In IXI, T1w, T2w, and Proton Density (PD)-weighted
images from 53 subjects were analyzed with a (25, 10, 18) split. In BRATS, T1-, T2-, and Fluid Attenuation Inversion Recovery (FR)-weighted images from 53 subjects were analyzed with a (25, 10, 20) split. In MIDAS, T1- and T2-weighted images from 66 subjects were analyzed with a (48, 5, 13) split. In OASIS, T1-, T2-, and FR-weighted images from 48 subjects were analyzed with a (22, 9, 17) split. Across the four datasets, the training set contained 2780, 2500, 3874, 2780 cross-sections per source-target configuration, respectively.

### B. Competing Methods

We demonstrated pFLSynth against a centralized benchmark [7] and federated baselines [47], [49], [58], [65], [66]. For each method, hyperparameter selection was performed via identical cross-validation procedures. All models shared generators across sites but used a separate local discriminator per site and per source-target configuration for improved performance.

**Central:** A non-federated synthesis model was considered following data aggregation across sites [7]. Loss function and architecture matched pFLSynth, albeit the mapper and PBs were excluded. The Central model serves as a privacy-violating benchmark for pFLSynth.

**FedGAN:** A federated synthesis model was implemented with matching loss function and architecture to pFLSynth, but without the mapper and PBs [58]. FedGAN aggregated the entire generator, so it serves as a non-personalized benchmark.

**FedMRI:** A federated model proposed for maintaining site specificity in MRI reconstruction was considered [65]. FedMRI used a UNet backbone with a shared encoder and site-specific decoders [65]. For adversarial synthesis, a discriminator with matching loss function to pFLSynth was used.

**FedMM:** Federated implementation of a unified synthesis model (MM-GAN) was considered [66]. FedMM used a UNet generator and multiple input-output channels to cope with different source-target configurations [66]. The entire generator was aggregated. For fair comparison, curriculum learning was omitted to ensure standard sample selection.

**FedMed:** A recent federated model for MRI synthesis was considered [49]. FedMed (abbreviated from FedMed-GAN) used a UNet generator that was entirely aggregated [49]. FedMed was originally proposed for unpaired synthesis. Only the forward mapping generator was retained for paired synthesis, and matching loss function to pFLSynth was used. For

![Fig. 3: Federated synthesis under a common task configuration within and across sites, T₁→T₂. Source, reference target, and synthetic target images from competing methods are shown. Representative results are displayed for a) IXI and b) BRATS. Overall, pFLSynth synthesizes images with fewer artifacts and lower noise levels compared to other federated methods.](image-url)

|                      | IXI         | BRATS       | MIDAS       | OASIS       |
|----------------------|-------------|-------------|-------------|-------------|
|                      | T₁→T₂       | T₁→T₂       | T₁→T₂       | T₁→T₂       |
| Central              | P↑ 28.6±1.3 | 26.1±0.9    | 28.1±0.5    | 25.2±0.6    |
|                      | S↑ 94.3±1.3 | 93.0±1.1    | 91.9±0.9    | 83.7±2.4    |
|                      | F↑ 7.4      | 24.8        | 9.7         | 18.1        |
| pFLSynth             | P↑ 28.6±1.3 | 26.3±1.0    | 28.2±0.5    | 25.1±0.6    |
|                      | S↑ 94.4±1.4 | 93.3±1.1    | 92.1±0.8    | 84.4±2.0    |
|                      | F↑ 7.7      | 23.9        | 9.2         | 30.2        |
| FedGAN               | P↑ 27.6±1.1 | 26.0±0.9    | 27.9±0.6    | 24.6±0.7    |
|                      | S↑ 93.2±1.3 | 92.9±1.1    | 91.6±0.6    | 83.4±2.6    |
|                      | F↑ 11.5     | 24.2        | 9.9         | 40.2        |
| FedMRI               | P↑ 27.9±1.2 | 25.7±0.8    | 27.7±0.5    | 23.8±0.8    |
|                      | S↑ 93.7±1.3 | 92.9±1.1    | 91.2±1.0    | 82.7±2.1    |
|                      | F↑ 9.6      | 32.5        | 11.3        | 43.0        |
| FedMM                | P↑ 26.9±1.1 | 25.6±0.8    | 27.6±0.6    | 24.3±0.5    |
|                      | S↑ 91.7±1.9 | 92.0±1.1    | 91.5±0.9    | 80.0±2.4    |
|                      | F↑ 18.1     | 28.3        | 9.5         | 42.3        |
| FedMed               | P↑ 27.1±1.1 | 25.6±0.7    | 27.8±0.6    | 24.3±0.5    |
|                      | S↑ 92.5±1.4 | 92.4±1.1    | 91.3±0.8    | 81.0±1.9    |
|                      | F↑ 13.0     | 26.6        | 9.4         | 47.0        |
| FedCycle             | P↑ 27.2±1.1 | 25.5±0.8    | 27.8±0.6    | 24.2±0.5    |
|                      | S↑ 92.8±1.5 | 92.4±1.1    | 91.5±0.9    | 80.8±2.2    |
|                      | F↑ 13.4     | 23.5        | 9.8         | 45.8        |

**TABLE I:** Performance in a common task within and across sites (i.e., T₁→T₂). A centrally trained benchmark is also reported. PSNR (P, dB) and SSIM (S, %) are listed as mean±std across test subjects, whereas FID (F) summarizes the entire test set. Boldface indicates the top-performing federated model for each site, task, and metric.
a) Source | Reference | pFLSynth | FedGAN | FedMRI | FedMM | FedMed | FedCycle

b) Source | Reference | pFLSynth | FedGAN | FedMRI | FedMM | FedMed | FedCycle

Fig. 4: Federated synthesis under a variable task configuration within and across sites (T₁→T₁ and T₂→T₂ in BRATS, T₁→T₂ and T₂→T₁ in MIDAS, T₂→T₁, and T₂→FR in OASIS). Source, reference target, and synthetic target images from competing methods are shown. Representative results are displayed for a) T₂→PD in IXI and b) T₂→FR in OASIS. Overall, pFLSynth synthesizes images with fewer artifacts and lower noise levels compared to other federated methods.

fair comparison, differential privacy procedures were omitted. FedCycle: Originally proposed for low-dose CT denoising, FedCycle (abbreviated from FedCycleGAN) is a federated model based on a switchable backbone to reduce model complexity [47]. FedCycle was implemented with a UNet generator and adaptive normalization as in [47]. To adopt FedCycle for paired synthesis, only the forward mapping generator was retained, and matching loss function to pFLSynth was used. The switching mechanism was used to adapt the model to different source-target configurations as in pFLSynth.

C. Architectural Details

M in pFLSynth had \( L_M = 6 \) layers, \( S \) followed an encoder-bottleneck-decoder structure with \( L_S = 15 \) stages. The encoder had 3 stages \((e1−e3)\), each with a convolutional layer followed by a PB, and kernel sizes were 7, 3, 3 across stages. The bottleneck had 9 residual layers \((r1−r9)\) with kernel size 3 [59], each residual layer followed by a PB. The decoder had 3 convolutional layers \((d1−d3)\) of kernel sizes 3, 3, 7, with the first two layers followed by PBs. \( D \) had 5 convolutional layers of kernel size 4. An FL setup with \( K = 4 \) different datasets was considered, so a site index of \( v^S ∈ \mathbb{Z}_2^4 \) was used. Given multiple source-task configurations, uniform random sample selection was utilized for learning different tasks. The datasets examined included T₁, T₂, PD and FR contrasts, so a source-target configuration index of \( u_c ∈ \mathbb{Z}_2^8 \) was formed. The mapper received these indices and produced a latent vector \( w^F ∈ \mathbb{R}^{512} \). The MLP in AdaCW had 2 layers with 64 hidden neurons. The splitting layer in the synthesizer \((L_U)\) was selected as 15 based on validation performance.

D. Modeling Procedures

For fair comparison, all competing models were trained using the same discriminator (Eq. 12) and compound synthesis loss (Eq. 17). Hyperparameters including pixel-wise loss weight, number of communication rounds, number of epochs, and learning rate were selected via cross-validation. A common set of hyperparameters that yielded near-optimal results across models and datasets were selected. Training was performed via the Adam optimizer with batch size of 1 and \( β_1 = 0.5, β_2 = 0.999 \). For centralized models, training lasted 150 epochs. Training lasted \( P=150 \) rounds for federated models with \( E=1 \) local epochs each. Learning rate was set as 0.0002 during the initial 75 epochs and linearly decayed to 0 during the remaining epochs. The pixel-wise loss weight was set to \( λ_{pix} = 100 \). Models were implemented using PyTorch and Nvidia RTX 3090 GPUs. Performance was evaluated via PSNR, Structural Similarity Index (SSIM), and Frechet Inception Distance (FID) [67] metrics between synthetic and reference target-contrast images. Results were reported as mean and standard deviation across test subjects. Wilcoxon signed-rank tests were conducted to assess the significance of PSNR/SSIM differences among competing methods.

V. Results

A. Common Task Configuration

FL experiments were conducted in a four-site setup based on IXI, BRATS, MIDAS, and OASIS datasets. We first demonstrated pFLSynth for decentralized modeling of a single synthesis task across sites (i.e., T₁→T₂). This configuration reflects the influence of implicit heterogeneity in the data distribution due to cross-site differences in scanners and protocols used to collect the same MRI contrasts. Performance metrics for competing methods are listed in Table I. pFLSynth outperforms all federated baselines at each site (\( p<0.05 \)), and offers on par performance with the central benchmark. On average across sites, pFLSynth achieves 0.8dB PSNR, 1.4% SSIM, 5.8 FID improvement over competing FL baselines. Representative images are displayed in Fig. 3.
reduces artifacts and noise, with the closest tissue depiction to the ground-truth target images, particularly near pathology.

### B. Variable Task Configuration

We then demonstrated pFLSynth for simultaneous modeling of diverse tasks within and across sites (T₁ → T₂ and T₂ → PD in IXI, T₁ → T₂ and FR → T₁ → T₂ in BRATS, T₁ → T₂ and T₂ → T₁ in MIDAS, T₁ → T₂ and T₂ → FR in OASIS). This variable configuration creates explicit heterogeneity rendering suboptimal performance in a unified synthesis model, even for the central benchmark. Yet, we hypothesized that the personalized pFLSynth model should cope more reliably with heterogeneity. Performance metrics are listed in Table II. pFLSynth outperforms all federated baselines at each site (p < 0.05), except for MIDAS where FedCycle has higher PSNR when par SSIM in T₁ → T₂, FedGAN has higher PSNR/SSIM in T₂ → T₁, FedMix has on par SSIM in T₁ → T₂, and for OASIS where FedMM has lower FID in T₁ → T₂. On average across tasks and sites, pFLSynth achieves 2.0dB PSNR, 5.0% SSIM, 24.9 FID improvement over federated baselines. pFLSynth also performs competitively against the central benchmark improving PSNR by 2.0dB, SSIM by 7.0%, FID by 18.5. Representative images are shown in Fig. 4. Again, pFLSynth reduces artifacts and noise to accurately depict brain tissue. Altogether, the findings on common and variable configurations indicate the robustness of pFLSynth against implicit and explicit heterogeneity in MRI synthesis.

### C. Ablation Studies

Ablation studies were conducted to assess the contributions of major design elements in pFLSynth. First, pFLSynth was compared against variants where AdaIN layers were ablated, AdaCW layers were ablated, full network aggregation was used, and the mapper was ablated (i.e., site and source-target configuration indices directly input to PBs). Performance metrics for the common task configuration are listed in Table III. pFLSynth outperforms all ablated variants at each site, except for the mapper-ablated variant that yields slightly lower FID in IXI, and the AdaIN-ablated variant that yields slightly lower FID in BRATS. These results demonstrate the importance of PBs, partial network aggregation, and the mapper to synthesis performance. We then compared pFLSynth against variants where site index was removed, and source-target configuration index was removed. Performance in the variable task configuration is summarized in Table IV. pFLSynth outperforms ablated variants across sites and tasks, except for occasionally albeit slightly higher PSNR and a few other metrics.
SSIM with variants in MIDAS. These results demonstrate the importance of using personalized latents per site and source-target configuration.

Next, we examined the effect of delayed insertion of a specific site or translation task to the FL setup. To do this, spare digits were reserved in the site index and source-target configuration index to code late joiners included halfway during the training. For delayed site analysis, a single site was held out in the common task configuration and pFLSynth was compared to a variant with ablated site index. For delayed task analysis, a single task was held out in the variable task configuration and pFLSynth was compared to a variant with ablated source-target configuration index. Performance metrics for the held-out site and task are listed in Table V (there were unsubstantial differences for non-held-out sites/tasks). Models with delayed site or task perform competitively with the original model including all sites and tasks, and they outperform variants with ablated site or source-target configuration index. These results suggest that pFLSynth shows a degree of inherent reliability against delayed insertion.

Lastly, we assessed the influence of PBs and PNA on possible information leakage among FL sites. Recent studies posit layer-wise measures to assess leakage in network models [68]. Thus, we measured the similarity of activation maps in local synthesizer layers to assess the potential for leakage [69]. A random set of 50 training source images were selected from each site and projected separately through all local synthesizers. Similarity for a given source image was taken as Spearman’s correlation coefficient between the activation maps it elicits in separate sites. Fig. 5 displays similarity for pFLSynth, variants with ablated PBs and/or PNA, and a variant that shared PBs across sites. Similarity of activation maps is lowered by inclusion of both PBs and PNA. Unshared PBs in pFLSynth further reduce similarity in activation maps, suggesting enhanced reliability against leakage.

D. Complexity and Communication Efficiency

A practical concern for FL models is efficiency in decentralized settings. For each model, Table VI lists number of total versus communicated parameters. While pFLSynth has moderately higher complexity, it has comparable training and inference times to other models (unreported). Importantly, as pFLSynth does not communicate upstream generator layers, it performs competitively in communication efficiency, showcasing an added benefit of PNA.

### VI. Discussion

Federated MRI synthesis has to operate under distributional heterogeneity in multi-site imaging data [21]. A recent study has considered FedAvg optimization of cycle-consistent models for MRI synthesis [49]. However, no prior study has proposed a dedicated mechanism to address data heterogeneity beyond the level inherently offered by FedAvg. To our knowledge, pFLSynth is the first FL method to personalize a global synthesis model to each individual site and translation task. Experiments on multi-site MRI data demonstrate that pFLSynth offers on par performance to a central benchmark while outperforming federated baselines. Our results suggest that pFLSynth can improve generalizability and flexibility in multi-site collaborations by permitting training on imaging

---

**TABLE IV:** Performance for pFLSynth and variant models ablated of site index ($v^k$) and source-target configuration index ($u_c$).

|       | IXI | BRATS | MIDAS | OASIS |
|-------|-----|-------|-------|-------|
|       | $T_1 \rightarrow T_2$ | $T_2 \rightarrow PD$ | $T_1 \rightarrow T_2$ | $T_2 \rightarrow T_1$ | $T_1 \rightarrow T_2$ | $T_2 \rightarrow FR$ |
| pFLSynth | $\uparrow$ | $27.9 \pm 1.2$ | $31.2 \pm 1.0$ | $25.6 \pm 1.0$ | $23.4 \pm 1.0$ | $26.0 \pm 0.6$ | $25.1 \pm 1.1$ | $23.4 \pm 1.8$ | $21.1 \pm 1.9$ |
| $S \downarrow$ | $93.7 \pm 1.5$ | $97.2 \pm 0.5$ | $92.4 \pm 1.2$ | $88.8 \pm 1.8$ | $89.0 \pm 1.2$ | $84.1 \pm 2.1$ | $81.7 \pm 3.3$ | $76.4 \pm 5.2$ |
| $F \downarrow$ | $8.6$ | $22.1$ | $23.2$ | $42.6$ | $10.4$ | $12.4$ | $46.7$ | $29.7$ |
| w/o $v^k$ | $\uparrow$ | $27.6 \pm 1.2$ | $30.9 \pm 1.0$ | $25.2 \pm 1.0$ | $18.5 \pm 1.0$ | $25.6 \pm 0.6$ | $23.8 \pm 1.0$ | $23.2 \pm 1.7$ | $19.7 \pm 1.9$ |
| $S \downarrow$ | $93.4 \pm 1.4$ | $97.0 \pm 0.4$ | $91.5 \pm 1.1$ | $80.5 \pm 2.0$ | $89.2 \pm 1.2$ | $82.2 \pm 1.2$ | $74.9 \pm 4.5$ | $71.1 \pm 6.3$ |
| $F \downarrow$ | $13.7$ | $22.5$ | $36.2$ | $147.9$ | $17.8$ | $25.0$ | $49.4$ | $39.7$ |
| w/o $u_c$ | $\uparrow$ | $27.4 \pm 1.2$ | $30.3 \pm 0.9$ | $24.6 \pm 1.1$ | $7.1 \pm 8.7$ | $25.8 \pm 0.6$ | $25.2 \pm 1.0$ | $22.2 \pm 1.6$ | $20.5 \pm 2.0$ |
| $S \downarrow$ | $93.6 \pm 1.4$ | $97.0 \pm 0.4$ | $91.2 \pm 1.2$ | $29.7 \pm 2.3$ | $88.9 \pm 1.1$ | $84.7 \pm 2.1$ | $80.5 \pm 3.1$ | $74.5 \pm 5.4$ |
| $F \downarrow$ | $9.8$ | $22.7$ | $23.5$ | $162.8$ | $10.6$ | $13.0$ | $47.3$ | $31.1$ |

**TABLE V:** Effects of delayed insertion. Results shown for the original pFLSynth with all sites/tasks included, pFLSynth with delayed site/task insertion, and variant models ablated of site or task index.

|       | IXI | BRATS | MIDAS | OASIS |
|-------|-----|-------|-------|-------|
|       | $T_1 \rightarrow T_2$ | $T_2 \rightarrow PD$ | $T_1 \rightarrow T_2$ | $T_2 \rightarrow T_1$ | $T_1 \rightarrow T_2$ | $T_2 \rightarrow FR$ |
| pFLSynth (original) | $\uparrow$ | $25.1 \pm 0.6$ | $84.4 \pm 2.0$ | $97.2 \pm 0.5$ | $31.2 \pm 1.0$ | $97.2 \pm 0.5$ | $22.1$ |
| $S \downarrow$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ |
| $F \downarrow$ | $30.2$ | $30.2$ | $30.2$ | $30.2$ | $30.2$ | $30.2$ | $30.2$ |
| pFLSynth (delayed site) | $\uparrow$ | $24.8 \pm 0.8$ | $82.9 \pm 2.2$ | $97.0 \pm 0.5$ | $30.9 \pm 1.0$ | $97.0 \pm 0.5$ | $22.1$ |
| $S \downarrow$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ |
| $F \downarrow$ | $3.6$ | $3.6$ | $3.6$ | $3.6$ | $3.6$ | $3.6$ | $3.6$ |
| w/o $v^k$ | $\uparrow$ | $23.0 \pm 1.1$ | $68.8 \pm 3.3$ | $96.6 \pm 0.5$ | $80.5 \pm 0.9$ | $96.6 \pm 0.5$ | $24.7$ |
| $S \downarrow$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ | $2.2 \pm 0.5$ |
| $F \downarrow$ | $44.6$ | $44.6$ | $44.6$ | $44.6$ | $44.6$ | $44.6$ | $44.6$ |

**Fig. 5:** Similarity of activation maps across sites in different stages of the synthesizer. A set of training images were projected through the local synthesizers at each site, and the resultant activation maps were compared across sites. Results are shown as mean±std of Spearman’s correlation coefficient across sites.
data from diverse sources and protocols.

FL avoids transfer of imaging data to mitigate patient privacy risks. Yet, inference attacks might leak information about training data from model parameters [8]. Here, we considered an FL setup where only generators were shared across sites. Yet, the discriminators were never communicated since our experiments in the initial phases of the study indicated that sharing discriminators did not elicit any performance benefit. This setup is reported to be relatively resilient against inference attacks [70]. Nevertheless, potential risks can be further minimized by adopting differentially private training [49], [71], or by extending the size and diversity of the training datasets to implicitly improve privacy [72]. Future studies are warranted to systematically examine the privacy properties of FL-based methods in multi-contrast MRI synthesis.

Here, we primarily considered an FL setup where all sites contributed to the entire training process. When late-joining sites were present, successful learning was achieved given a reasonable number of communication rounds after inclusion. An alternative is to use local mappers at each site to learn site-specific latents without needing a site index. Certain scenarios might involve inference at a site that was completely excluded from training. In those cases, an average model across training sites might yield suboptimal performance at a held-out site. To improve performance, transfer-learning procedures can be adopted to fine-tune the trained model on a compact dataset from the held-out site [7].

The proposed method can be developed along several technical lines. We performed training on paired datasets with registered source and target images from a matching set of tasks involving CT or PET, and other dense-prediction tasks such as reconstruction or super-resolution.

VII. Conclusion

We introduced a novel personalized FL method for multi-contrast MRI synthesis based on an adversarial model equipped with personalization blocks and partial network aggregation. Benefits over prior federated methods were demonstrated for brain image synthesis considering different public datasets and source-target configurations. Improved generalization against implicit and explicit domain shifts renders pFLSynth a promising candidate for multi-site collaborations. pFLSynth might also be adopted for other image translation tasks involving CT or PET, and other dense-prediction tasks such as reconstruction or super-resolution.

REFERENCES

[1] S. Bakas et al., “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features,” Sci. Data, vol. 4, 2017.
[2] K. Krupa and M. Bekesińska-Figatowska, “Artifacts in Magnetic Resonance Imaging,” Pol. J. Radiol., vol. 80, pp. 93–106, 2015.
[3] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Gloker, K. Van Leemput, and B. Fischl, “Is Synthesizing MRI Contrast Useful for Inter-modality Analysis?” in Med. Image Comput. Comput. Assist. Interv., 2013, pp. 1013–1023.
[4] V. Sevetlidić, M. V. Giaffrida, and S. A. Tsaftaris, “Whole Image Synthesis Using a Deep Encoder-Decoder Network,” in Simul. Synth. Med. Imaging, 2016, pp. 127–137.
[5] T. Joyce, A. Chartsias, and S. A. Tsaftaris, “Robust multi-modal MR image synthesis,” in Med. Image Comput. Comput. Assist. Interv., 2017, pp. 347–355.
[6] W. Wei et al., “Fluid-attenuated inversion recovery MRI synthesis from multisequence MRI using three-dimensional fully convolutional networks for multiple sclerosis,” J. Med. Imaging, vol. 6, no. 1, p. 014005, 2019.
[7] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, “Image Synthesis in Multi-Contrast MRI With Conditional Generative Adversarial Networks,” IEEE Trans. Med. Imag., vol. 38, no. 10, pp. 2375–2388, 2019.
[8] G. A. Kaissis, M. R. Makowski, D. Rueckert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” Nat. Mach. Intell., vol. 2, no. 6, pp. 305–311, 2020.
[9] W. Li et al., “Privacy-Preserving Federated Brain Tumour Segmentation,” in Mach. Learn. Imag., 2019, pp. 133–141.
[10] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, “Multi-institutional Deep Learning Model Training Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation,” in Med. Image Comput. Comput. Assist. Interv., 2019, pp. 92–104.
[11] N. Rieke et al., “The future of digital health with federated learning,” npj Digital Medicine, vol. 3, no. 1, p. 119, 2020.
[12] H. R. Roth et al., “Federated Learning for Breast Density Classification: A Real-World Implementation,” in DART, DCL, 2020, p. 181–191.
[13] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, “Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results,” Med. Image. Anal., vol. 65, p. 101765, 2020.
[14] Q. Liu, C. Chen, J. Qin, Q. Dou, and P. Heng, “FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space,” in Comput. Vis. Pattern Recognit., 2021, pp. 1013–1023.
[15] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in AISTATS, 2017.
[16] J. Wang et al., “A Field Guide to Federated Optimization,” arXiv:2107.06917, 2021.
[17] T. Li, A. K. Sahu, A. S. Talwalkar, and V. Smith, “Federated Learning: Challenges, Methods, and Future Directions,” IEEE Signal. Process. Mag., vol. 37, pp. 50–60, 2020.
[18] N. Rieke et al., “The future of digital health with federated learning,” npj Digital Medicine, vol. 3, 12 2020.
[19] M. Sheller et al., “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data,” Sci. Rep., vol. 10, 07 2020.
[20] H. R. Roth et al., “Federated Whole Prostate Segmentation in MRI with Personalized Neural Architectures,” in Med. Image Comput. Comput. Assist. Interv., Cham: Springer, 2021, pp. 357–366.
[21] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, “FedBN: Federated Learning on Non-IID Features via Local Batch Normalization,” in Int. Conf. Learn. Represent., 2021.
[22] C. I. Bercea, B. Wiestler, D. Rueckert, and S. Albarqouni, “FedDis: Disentangled Federated Learning for Unsupervised Brain Pathology Segmentation,” arXiv:2103.03705, 2021.
[23] S. Pati et al., “The Federated Tumor Segmentation (FeTS) Challenge,” arXiv:2105.05874, 2021.

TABLE VI: Local model complexity (millions of parameters) and communication load (millions of parameters) for competing methods.
