Medical Image Analysis 94 (2024) 103121

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

One model to unite them all: Personalized federated learning of multi-contrast MRI synthesis

Onat Dalmaz [a,b], Muhammad U. Mirza [a,b], Gokberk Elmas [a,b], Muzaffer Ozbey [a,b], Salman U.H. Dar [a,b], Emir Ceyani [c], Kader K. Oguz [d], Salman Avestimehr [c], Tolga Çukur [a,b,e,*]

[a] Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey
[b] National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara 06800, Turkey
[c] Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA
[d] Department of Radiology, University of California, Davis Medical Center, Sacramento, CA 95817, USA
[e] Neuroscience Program, Bilkent University, Ankara 06800, Turkey

ARTICLE INFO

Keywords:
Federated learning
Personalization
Heterogeneity
MRI
Synthesis
Translation

ABSTRACT

Curation of large, diverse MRI datasets via multi-institutional collaborations can help improve learning of generalizable synthesis models that reliably translate source- onto target-contrast images. To facilitate collaborations, federated learning (FL) adopts decentralized model training while mitigating privacy concerns by avoiding sharing of imaging data. However, conventional FL methods can be impaired by the inherent heterogeneity in the data distribution, with domain shifts evident within and across imaging sites. Here we introduce the first personalized FL method for MRI Synthesis (pFLSynth) that improves reliability against data heterogeneity via model specialization to individual sites and synthesis tasks (i.e., source-target contrasts). To do this, pFLSynth leverages an adversarial model equipped with novel personalization blocks that control the statistics of generated feature maps across the spatial/channel dimensions, given latent variables specific to sites and tasks. To further promote communication efficiency and site specialization, partial network aggregation is employed over later generator stages while earlier generator stages and the discriminator are trained locally. As such, pFLSynth enables multi-task training of multi-site synthesis models with high generalization performance across sites and tasks. Comprehensive experiments demonstrate the superior performance and reliability of pFLSynth in MRI synthesis against prior federated methods.

1. Introduction

MRI synthesis

MRI can provide non-invasive anatomical assessment with rich diagnostic information accumulated over multiple tissue contrasts, albeit it suffers from prolonged scan times (Bakas et al., 2017). Costs associated with multi-contrast protocols often prohibit comprehensive acquisitions or repeat runs of corrupted contrasts during an exam, hampering clinical utilization (Krupa and Bekiesińska-Figatowska, 2015). MRI synthesis is a promising solution wherein unavailable target-contrast images of an anatomy are imputed from a subset of acquired source-contrast images (Iglesias et al., 2013; Van Nguyen et al., 2015; Vemulapalli et al., 2015; Jog et al., 2017). Naturally, with separate MRI contrasts capturing partially distinct information about tissue properties, source versus target contrasts must be carefully chosen to ensure that the resultant synthesis task is reasonably well-posed (Lee et al., 2019). Consequently, an important clinical use scenario for MRI synthesis is to impute the subset of sequences within a comprehensive imaging protocol that capture relatively redundant information in the context of diagnosing specific diseases. Literature suggests that individual MRI sequences show varying degrees of sensitivity to underlying pathology depending on disease (Atlas, 2009; Ellison et al., 2012). For instance, T_1-weighted images are considered relatively more useful than T_2-weighted images in evaluation and analysis of cerebral atrophy induced by Alzheimer's Disease, albeit T_2-weighted images are relatively more sensitive than T_1-weighted images to the overall number and distribution of cerebral lesions in Multiple Sclerosis (Adam et al., 2014). While T_2- and FLAIR-weighted images are commonly adopted in assessments of lesions with increased tissue fluid such as tumors, FLAIR-weighted images are suggested to be more sensitive than T_2-weighted images in detection of peritumoral edema (Adam et al., 2014). Thus, disease-specific prioritization can be exercised by acquiring pathology-sensitive sequences as sources and imputing remaining sequences as targets. In

[*] Corresponding author at: Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey.
E-mail address: cukur@ee.bilkent.edu.tr (T. Çukur).

https://doi.org/10.1016/j.media.2024.103121
Received 26 May 2023; Received in revised form 20 February 2024; Accepted 21 February 2024
Available online 23 February 2024

this way, MRI synthesis can improve time- and cost-efficiency of exams at busy imaging sites.

Another important use scenario is to facilitate participation in clinical studies to improve quality of downstream analyses. For instance, retrospective studies analyze pre-existing imaging data originally collected for different purposes, so there can be notable variations in acquired sequences across patients (Halligan et al., 2020). Yet, because patients might be unavailable and their tissue attributes might be altered due to disease-related evolution over the elapsed time between the original MRI exams and the study, retrospective designs do not permit later acquisition of missing sequences (Singh et al., 2022). In turn, this can restrict the scope of the study to a smaller population for whom the desired imaging data are available, or to a limited imaging protocol that is commonly available across a broad population (Altman and Bland, 2007). Another example is drug investigations that typically seek to lower bias by maintaining high diversity in the trial population, including genetic diversity that might necessitate collaboration among geographically-distant sites (Rodney, 2021). However, many candidate sites might be unable to run comprehensive imaging protocols on their local population due to resource limitations, rendering them ineligible for participation (Clark et al., 2019). In such cases, protocol consistency can be enhanced by imputing missing sequences from the subset of acquired sequences in each patient. In this way, MRI synthesis might increase the number of participants eligible for enrollment in a clinical study when a demanding multi-contrast MRI protocol is desired.

Centralized learning for MRI synthesis

In recent years, learning-based synthesis models locally trained on data from an individual imaging site have offered leaps in translation performance (Sevetlidis et al., 2016; Joyce et al., 2017). Yet, such single-site models show poor generalization to features unencountered during training, so they typically underperform when tested on separate sites due to native domain shifts in image features (Wei et al., 2019; Dar et al., 2019). Among recent approaches to improve generalization in medical imaging, a contrast-agnostic method has been proposed based on model training with synthetic data obtained via physics-driven image generation at randomized contrasts and resolutions (Billot et al., 2023a,b). While such randomization was reported to improve robustness in segmentation models, domain shifts in synthesis tasks include non-linear variations in the MRI signal due to interactions between sequence and tissue-relaxation parameters (Denck et al., 2021). Thus, it may be non-trivial to devise physics-driven generation for synthesis tasks without knowledge of underlying relaxation parameters (Wang et al., 2020a), and the utility of synthetic training data remains to be demonstrated for MRI synthesis models.

Another important approach to improve generalization is to centrally train learning-based models on a large and diverse set of actual imaging data. One strategy is to curate a broad dataset within a single site with ample resources, including access to multiple scanners (Wasserthal et al., 2023). While centralized segmentation models trained on such single-site CT datasets were shown to offer reliable generalization across scanners (Wasserthal et al., 2023), the greater inter-scanner variability in multi-contrast MRI data can limit performance (Knoll et al., 2020). Moreover, geographic restrictions on recruited subjects can limit diversity in single-site datasets. Thus, the utility of this approach in building reliable MRI synthesis models remains to be demonstrated. An alternative strategy is to establish a multi-institutional collaboration to curate a broad dataset by aggregating data acquired at separate sites (Huang et al., 2018). Although it can natively enhance data diversity, this centralized approach can introduce patient privacy risks during transfer of sensitive imaging data (Kaissis et al., 2020). Since modeling is performed centrally by processing the broad multi-site dataset, it also requires access to significant compute resources at the site(s) that eventually build models.

Federated learning for MRI synthesis

Federated learning (FL) is a recent framework for multi-institutional collaborations based on decentralized model training (Li et al., 2019a; Sheller et al., 2019; Roth et al., 2021). Over multiple communication rounds, an FL server sporadically aggregates locally-trained models at each site to compute a global multi-site model (McMahan et al., 2017), thereby enabling collaborative learning on multi-site datasets. FL can offer several potential benefits over centralized-learning approaches. First, FL alleviates patient privacy risks by transferring models instead of imaging data (Li et al., 2020b). Second, FL distributes costs associated with model training across multiple sites to lower demand for compute resources in individual sites (Kaissis et al., 2020). This could facilitate participation of relatively smaller sites with limited resources in collaborations. The ability to scale up multi-site collaborations without increasing privacy risks or compute demand per site might also help expand the scope of clinical studies on rare diseases, which often require large-scale collaborations to obtain sufficiently-large training datasets to build deep-learning models given low disease prevalence (Rieke et al., 2020).

A recent study suggests that multi-site synthesis models trained via FL promise improved generalization over single-site models locally trained on relatively limited and homogeneous datasets (Xie et al., 2022a). Yet, conventional FL methods are susceptible to data heterogeneity, resulting from differences in scan protocols and hardware within or across sites (Rieke et al., 2020; Sheller et al., 2020). Significant heterogeneity in multi-site datasets causes divergent statistics across spatial and channel dimensions of feature maps within learning-based models (Elmas et al., 2022). In turn, this can reduce the sensitivity for site-specific features and compromise model performance (Guo et al., 2021b). Previous FL studies on medical imaging have introduced several prominent approaches to cope with data heterogeneity in segmentation (Roth et al., 2021), classification (Li et al., 2021; Yan et al., 2021), and reconstruction (Guo et al., 2021b; Feng et al., 2021b; Elmas et al., 2022) tasks. However, to our knowledge, no prior study has addressed data heterogeneity in federated multi-contrast MRI synthesis.

In an idealized setup for federated MRI synthesis, all sites can collaborate to learn a shared set of translation tasks based on the commonly encountered varieties of acquired and missing contrasts (e.g., $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_1$ in all sites, source→target). In this setup, a multi-site model can benefit from training on diverse anatomy across different subject pools and from training on a larger set of image samples from a common source-target configuration. Yet, despite the task commonality, the resultant multi-site model can be susceptible to implicit data heterogeneity due to inter-site variations in sequence parameters and scanner hardware as well as intra-site task variations. Note that individual sites might also be interested in learning distinct translation tasks as they often prioritize different sequences in multi-contrast protocols (Sharma and Hamarneh, 2020). In this multi-task training setup, sites can collaborate while the acquired and missing contrasts are partially overlapping (e.g., $T_1 \rightarrow T_2$ in Site A, $T_2 \rightarrow$PD in Site B, $T_1 \rightarrow$PD in Site C). In this case, a multi-site model can still benefit from training on diverse anatomy across separate subject pools and from training on a larger dataset with overlapping contrasts. However, utilization of variable translation tasks will inevitably induce explicit data heterogeneity due to inter-site differences in source-target configurations. In turn, both implicit and explicit heterogeneity can cause notable performance degradation in regular synthesis models that lack personalization[1] mechanisms to maintain specialization (Dar et al., 2019).

[1] In the FL literature, individual clients are taken to be agents performing local learning of various tasks on their local data silos, and personalization is a common nomenclature for model specialization to the tasks performed by individual clients (Fallah et al., 2020; Mansour et al., 2020; Tan et al., 2023; Roth et al., 2021). In the context of MRI synthesis, each client can be

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 improve model specialization, novel personalization blocks are introduced that receive these latent variables to control the statistics of generated feature maps. To improve communication efficiency and personalization, we further propose partial network aggregation on later generator stages, while earlier generator stages and discriminator stages are kept local. These design elements enable pFLSynth to reliably generalize across multiple sites and diverse synthesis tasks. Comprehensive experiments on multi-site MRI data clearly demonstrate the superior performance of pFLSynth against prior federated models. Codes to implement pFLSynth and baselines examined in this study are 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.
- pFLSynth leverages novel personalization blocks that improve specialization to individual sites and source-target configurations by using site- and task-specific latent variables to modulate feature map statistics across spatial and channel dimensions.
- To improve communication efficiency and personalization, partial network aggregation is adopted on later generator stages while earlier generator stages and the discriminator are kept site-specific.

2. Related work

2.1. Centralized deep-learning models

Performance in multi-contrast MRI synthesis tasks has witnessed a recent surge with adoption of deep-learning models (Bowles et al., 2016; Chartsias et al., 2017; Nie et al., 2018). In the traditional learning framework, multi-contrast MRI data from one or more sites are first curated to a central dataset, transferring imaging data across sites as necessary. A synthesis model is then centrally trained to translate source onto target images on the curated dataset. Among numerous models, earlier studies proposed convolutional neural networks (CNN) with pixel-wise loss terms (Sevetlidis et al., 2016; Joyce et al., 2017; Wei et al., 2019; Cordier et al., 2016; Zhang et al., 2019). To improve capture of tissue details, generative adversarial networks (GAN) were later introduced based on adversarial loss terms that indirectly learn the distribution of target images (Armanious et al., 2020; Beers et al., 2018; Dar et al., 2019; Lan et al., 2021; Yu et al., 2019; Zhou et al., 2020; Luo et al., 2021; Zhan et al., 2021; Yang et al., 2021). Commonly, maximal performance has been aimed by training a singular model for each separate source-target configuration, which can be experimentally burdening (Yurt et al., 2021). To improve practicality, some studies have instead proposed task-unified models capable of performing multiple translation tasks (Lei et al., 2020; Sharma and Hamarneh, 2020; Lee et al., 2019; Li et al., 2019b; Wang et al., 2020a; Dalmaz et al., 2022; Liu et al., 2023).

Despite their demonstrated success, studies following the traditional framework build either single-site models using a local dataset acquired at a given site (Guo et al., 2021a), or centralized models trained on a server that stores an aggregate dataset acquired from multiple sites (Yurt et al., 2021). Single-site models are commonly trained on relatively compact datasets given practical limitations on executing multi-contrast MRI exams on broad patient cohorts. As such, they can suffer from suboptimal learning due to limited number of training samples, and from poor generalization due to limited diversity in training data. While centralized models can help mitigate these limitations, they involve transfer of imaging data from separate sites onto a central server prior to modeling. This elevates privacy risks as imaging data contain sensitive information regarding patients' anatomy and health (Kaissis et al., 2020).

2.2. Federated deep-learning models

To remedy privacy concerns, the FL framework conducts decentralized training at multiple sites by communicating model parameters instead of data (Li et al., 2020b). The utility of FL methods have been demonstrated in several imaging tasks including segmentation (Sheller et al., 2019; Li et al., 2019a; Roth et al., 2021; Liu et al., 2021), classification (Yan et al., 2021; Park et al., 2021), reconstruction (Guo et al., 2021b; Song and Ye, 2021; Li et al., 2021, 2020a), and unconditional image generation (Che et al., 2022). Yet, the potential of decentralized procedures in medical image translation tasks remains relatively unexplored. Only few recent studies have reported multi-contrast MRI synthesis with FL based on cycle-consistent models (Xie et al., 2022a,b). These reports examined FL performance under approximately IID settings by partitioning a single dataset to emulate multiple sites, and building singular models for each separate source-target configuration. IID settings do not capture the scope of data heterogeneity expected in practical scenarios, and no dedicated personalization procedures were considered in these studies to cope with heterogeneity.

Here, we introduce a novel MRI synthesis method, pFLSynth, that leverages personalization mechanisms to improve reliability against data heterogeneity. To our knowledge, pFLSynth is the first personalized FL method for multi-contrast MRI synthesis that can offer reliable performance under non-IID settings. With similar aims to pFLSynth, several recent FL studies on other imaging tasks have aimed to address data heterogeneity. For classification tasks, domain adaptation based on a mixture-of-experts approach was proposed where global site-general and local site-specific models are combined (Li et al., 2020a). For classification and reconstruction, domain adaptation based on adversarial alignment among separate sites was proposed such that a common representational space is learned across sites (Yan et al., 2021; Li et al., 2020a; Liu et al., 2021; Guo et al., 2021b). While these domain-adaptation methods help alleviate across-site heterogeneity, they do not consider within-site heterogeneity due to task variability. Furthermore, they increase communication load for FL training since a set of images and/or latent representations are additionally transferred among sites for adaptation. In contrast, pFLSynth addresses both across- and within-site heterogeneity via novel personalization blocks that use site- and task-specific latent variables to alter the statistics of feature maps, and it does not communicate additional images or latent representations.

For classification, local batch normalization (BN) layers were introduced in a global network architecture (Li et al., 2021). Site specialization can be improved by training the normalization parameters in BN independently for each site. However, the use of fixed BN parameters within sites neglects within-site heterogeneity due to task variability, and learning BN parameters across image batches can reduce sensitivity to sample-specific features (Wu and He, 2018). Note also that multi-contrast MRI datasets manifest significant heterogeneity that can elicit divergent statistics in not only spatial but also channel dimensions of feature maps (Elmas et al., 2022; Wang et al., 2020b). To address these issues, pFLSynth leverages sample-specific adjustment of feature map statistics geared for individual sites and tasks to simultaneously cope with across- and within-site heterogeneity. Unlike batch-wise adjustments in BN layers, these sample-specific adjustments are mediated by

taken as a single imaging site, where a site denotes a group of researchers who have access to a specific MRI dataset and who agree to participate in an FL experiment. Each source-target contrast configuration would constitute a separate task. Thus, here we adopted the term "personalization" to describe model specialization to individual sites/tasks.

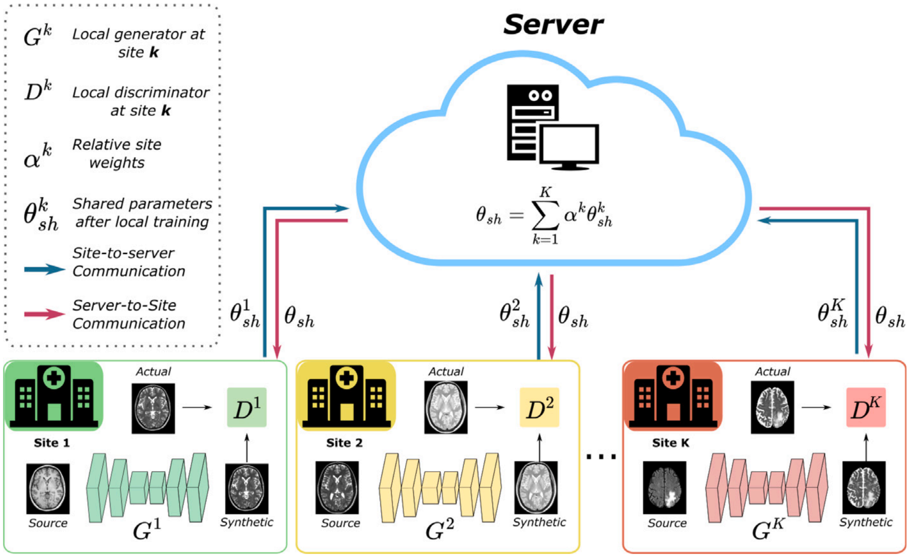

G^k Local generator at site **k**
D^k Local discriminator at site **k**
α^k Relative site weights
θ^k_{sh} Shared parameters after local training
Site-to-server Communication
Server-to-Site Communication
Server
$\theta_{sh} = \sum_{k=1}^{K} \alpha^k \theta^k_{sh}$
θ^1_{sh}
θ_{sh}
θ^2_{sh}
θ_{sh}
θ^K_{sh}
θ_{sh}
Site 1
Actual
D^1
Source
G^1
Synthetic
Site 2
Actual
D^2
Source
G^2
Synthetic
...
Site K
Actual
D^K
Source
G^K
Synthetic

Fig. 1. pFLSynth is a personalized federated learning (FL) method for multi-contrast MRI synthesis, based on an adversarial model with a generator (G) and a discriminator (D). The generator synthesizes a target-contrast image given as input a source-contrast image, whereas the discriminator aims to distinguish between actual and synthetic target-contrast images. The model is decentrally trained on data from K sites through multiple communication rounds between an FL server and individual sites. In each round, individual sites receive global model parameters (θ_{sh}) from the server, train local copies to maximize synthesis performance on their local data, and send updated local models (θ^k_{sh}) to the server. The server aggregates the local models into the global model with relative site weights (α^k).

personalization blocks equipped with site-task instance normalization and site-task channel attention subblocks.

For image reconstruction and classification, network splitting methods were proposed where the encoder segments of the network are shared across sites while the decoder segments are unshared (Feng et al., 2021b). This partial aggregation approach can elevate site specialization by tailoring the decoder to individual sites, and avoid communication of decoder parameters. That said, in the current study, we have observed that later network stages show stronger correlations across sites than earlier stages for MRI synthesis tasks. Accordingly, we devise pFLSynth to share later stages while keeping earlier stages local to focus specialization in weakly-correlated network segments. This partial network aggregation approach promotes earlier generator stages to transform source images from the site-specific image space towards a relatively site-independent latent space, and later generator stages to transform the latent representations from the site-independent latent space onto a contrast-specific image space. Taken together, these unique aspects enable pFLSynth to effectively address heterogeneity in multi-contrast data for reliable MRI synthesis.

3. Methods

3.1. MRI synthesis with adversarial models

Adversarial models have become pervasive in MRI synthesis due to their sensitivity for high-frequency features (Dar et al., 2019; Beers et al., 2018; Lee et al., 2019; Armanious et al., 2020). 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:

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

where $\mathbb{E}$ denotes expectation, $\mathcal{D}$ 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 λ_{pix}. The traditional framework trains a model centrally following aggregation of multi-site data in a central repository (Kaissis et al., 2020).

Alternatively, decentralized training can be performed via communication between an FL server hosting a global generator (G with θ_G), and sites keeping local copies (G^k for site $k \in \{1,\ldots,K\}$). Discriminators can be unshared across sites to minimize risk of information leak (Rasouli et al., 2020). In each communication round, local copies are initialized with the global model transmitted by the server ($\theta^k_G \leftarrow \theta_G$). Local models are then trained to minimize a local synthesis loss:

$$(\theta^k_G, \theta^k_D) = \underset{\theta^k}{\operatorname{argmin}} \; \mathcal{L}^k_{syn}(\mathcal{D}^k, \theta^k), \qquad (2)$$

where $\mathcal{D}^k$ are training data $(x^k_{s_c}, x^k_{t_c})$, and (s_c, t_c) denotes the cth source-target configuration at site k ($c \in \{1,\ldots,C\}$). After each round, local models are aggregated on the server via federated averaging (FedAvg; McMahan et al. (2017)):

$$\theta_G = \sum_{k=1}^{K} \alpha^k \theta^k_G. \qquad (3)$$

α^k denote relative site weights typically set to $\frac{n^k}{n}$, 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_{θ^*}) is eventually used for local inference:

$$\hat{x}^k_{t_c} = G^*(x^k_{s_c}). \qquad (4)$$

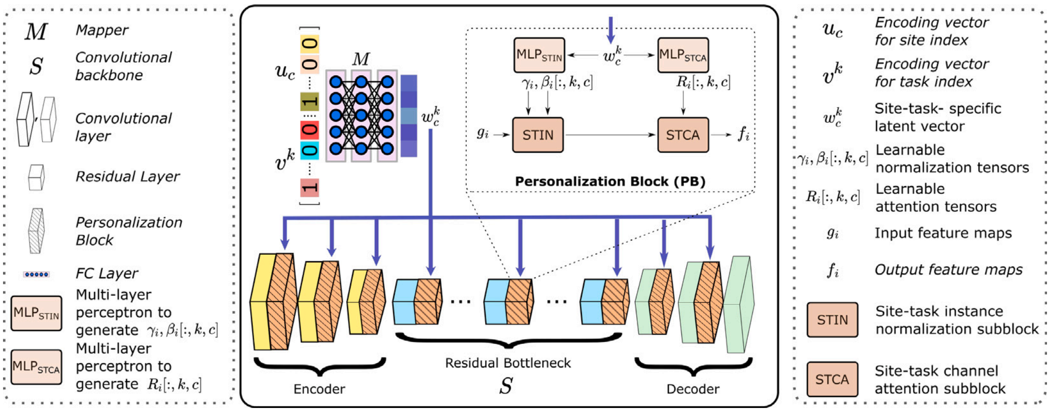
In this conventional FL approach, domain shifts due to data heterogeneity will compromise sensitivity of the global multi-site model to site- and task-specific features.

3.2. Personalized federated learning of MRI synthesis

Here we propose a personalized FL method for MRI synthesis, pFLSynth (Fig. 1). pFLSynth employs an adversarial model for synthesizing target images given as input source images along with site and source-target configuration information. The generator comprises a mapper that produces site- and task-specific latent variables, and a convolutional backbone equipped with personalization blocks (PBs) to map the source onto the target image (Fig. 2). PBs are composed of site-task instance normalization (STIN) and site-task channel attention

Table 1
Description of important variables related to multi-contrast MRI synthesis tasks and network components in pFLSynth.

Task-related variables	
k	Site index in {1, …,K}
$\mathcal{D}^k$	Local training data at site k
config^k	Set of source-target contrast configurations at site k
c	Source-target contrast configuration index in {1, …,C}
(s_c, t_c)	Source and target contrast pair for the cth configuration
$x^k_{s_c}$	Actual source image of contrast s_c at site k
$x^k_{t_c}$	Actual target image of contrast t_c at site k
$\hat{x}^k_{t_c}$	Synthetic target image of contrast t_c at site k
v^k	One-hot encoding vector for site index
u_c	One-hot encoding vector for source-target configuration index
w^k_c	Site- and task-specific latent variable vector
Network-related variables	
G^k	Local generator subnetwork at site k with params. θ^k_G
D^k	Local discriminator subnetwork at site k with params. θ^k_D
M^k	Mapper module of G^k with. params. θ^k_M
L_{cut}	Cut-point stage of the convolutional backbone in G^k
S^k_B	Later stages of the convolutional backbone in G^k with. params. $\theta^k_{S_B}$
S^k_A	Earlier stages of the convolutional backbone in G^k with. params. $\theta^k_{S_A}$
M	Global mapper module with. params. θ_M
S_B	Global, later stages of the convolutional backbone with. params. θ_{S_B}
CB_i	Convolutional block at the ith generator stage
PB_i	Personalization block at the ith generator stage
STIN	Site-task instance normalization subblock with params. β_i (mean), γ_i (std)
STCA	Site-task channel attention subblock with params. R_i (attention)



M
Mapper
S
Convolutional backbone
Convolutional layer
Residual Layer
Personalization Block
FC Layer
MLP_STIN
Multi-layer perceptron to generate γ_i, β_i[:, k, c]
MLP_STCA
Multi-layer perceptron to generate R_i[:, k, c]
u_c
v^k
0
0
1
0
0
1
M
w_c^k
MLP_STIN
MLP_STCA
γ_i, β_i[:, k, c]
R_i[:, k, c]
g_i
STIN
STCA
f_i
Personalization Block (PB)
Encoder
Residual Bottleneck
S
Decoder
u_c
Encoding vector for site index
v^k
Encoding vector for task index
w_c^k
Site-task- specific latent vector
γ_i, β_i[:, k, c]
Learnable normalization tensors
R_i[:, k, c]
Learnable attention tensors
g_i
Input feature maps
f_i
Output feature maps
STIN
Site-task instance normalization subblock
STCA
Site-task channel attention subblock

Fig. 2. pFLSynth's generator contains a mapper M based on a multi-layer perceptron (MLP) that produces site- and task-specific latent variables (w^k_c) given site identity index (v^k) and source-target configuration index (u_c). It also includes a convolutional backbone S with encoder, bottleneck and decoder stages to map source images onto target images. In each stage of S, a convolutional block (CB) that filters input feature maps is followed by a personalization block (PB), except for the final decoder stage. PBs receive w^k_c and input feature maps g_i to generate output feature maps f_i with modulated statistics. Each PB contains a site-task instance normalization (STIN) subblock that computes mean and std normalization parameters given w^k_c via $\mathrm{MLP}_{\mathrm{STIN}}$, and a site-task channel attention (STCA) subblock that computes attention parameters given w^k_c via $\mathrm{MLP}_{\mathrm{STCA}}$.

(STCA) subblocks that modulate the statistics of feature maps. During training, partial network aggregation (PNA) is adopted to improve site specialization and communication efficiency (Fig. 3). Important variables related to multi-site datasets and network components in pFLSynth are summarized in Table 1.

3.2.1. Network architecture

Generator (G): The mapper M is an L_M-layer multi-layer perceptron (MLP). Receiving a binary vector for site index ($v^k \in \mathbb{Z}^K_2$, $\mathbb{Z}_2 = \{0, 1\}$) and a binary vector for indices of source and target contrasts ($u_c \in \mathbb{Z}^{2C}_2$), it produces a latent variable vector:

$$w^k_c = M(v^k \oplus u_c), \tag{5}$$

where $w^k_c \in \mathbb{R}^J$, $\oplus$ is concatenation, J is vector dimensionality. Parameterized with θ_M, M produces site- and task-specific latent variables to drive PBs.

The convolutional backbone S receives as input a source image x_{s_c} and latent variable vector w^k_c to generate a target image $\hat{x}_{t_c}$. S is inspired by ResNet (Dar et al., 2019; He et al., 2016) with a residual bottleneck between an encoder and a decoder (see Fig. 2). Let $\{S_1, S_2, \dots, S_{L_S}\}$ with parameters $\theta_{S_{1,\dots,L_S}}$ be the set of generator stages. At the ith 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 first processed via CB_i:

$$g_i = \mathrm{CB}_i(f_{i-1}) \in \mathbb{R}^{F_i, H_i, W_i}, \tag{6}$$

where CB_i denotes the ith convolutional block comprising a cascade of a convolutional layer, a batch normalization layer and a nonlinear activation function.

MR images of a given anatomy acquired under implicit differences in imaging parameters at separate sites, or under explicit differences in sequences for separate tasks will show non-linear variations in relative tissue-signal levels. As such, for each site and each task, feature maps within S can elicit divergent intensity statistics across spatial and channel dimensions (Elmas et al., 2022). To mitigate these heterogeneities, we introduce novel PBs inserted after each convolutional block except for the final generator stage. Each PB receives site- and task-specific latent variables for learnable modulation of the mean and standard

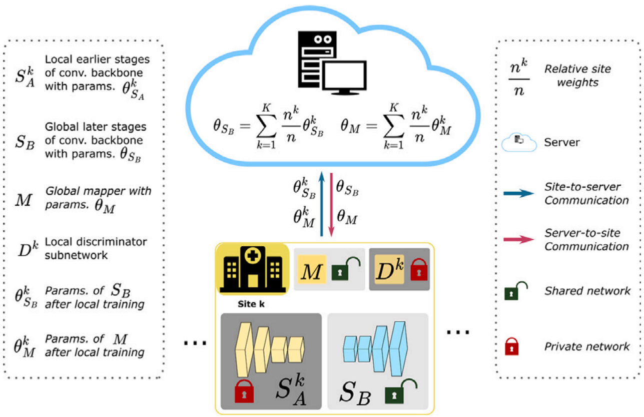

S_A^k Local earlier stages of conv. backbone with params. $\theta^k_{S_A}$
S_B Global later stages of conv. backbone with params. θ_{S_B}
M Global mapper with params. θ_M
D^k Local discriminator subnetwork
$\theta^k_{S_B}$ Params. of S_B after local training
θ^k_M Params. of M after local training
$\theta_{S_B} = \sum_{k=1}^{K} \frac{n^k}{n} \theta^k_{S_B}$ $\theta_M = \sum_{k=1}^{K} \frac{n^k}{n} \theta^k_M$
$\theta^k_{S_B}$ θ_{S_B}
θ^k_M θ_M
Site k
M
D^k
S_A^k
S_B
$\frac{n^k}{n}$ Relative site weights
Server
Site-to-server Communication
Server-to-site Communication
Shared network
Private network

Fig. 3. In conventional FL, locally-trained copies of the entire network are forwarded to the FL server for full network aggregation into a global model. In contrast, pFLSynth leverages partial network aggregation (PNA) to improve site specialization while lowering communication load. Prior to local training, a given site k receives the global mapper M and later stages of the generator S_B from the server. To compose a local synthesis model, these network components are combined with earlier stages of the generator S_A^k and the discriminator D^k that are maintained locally. The synthesis model is trained on local data, and then the updated local copies of S_B and M are sent to the server for aggregation.

deviation (std) of feature maps across S. This mechanism allows adaptation of the synthesis model to different signal-level distributions encountered at separate sites and for separate tasks.

In PB$_i$, an STIN subblock first modulates feature maps across spatial dimensions via instance normalization (Huang and Belongie, 2017), according to learnable normalization tensors γ_i and $\beta_i \in \mathbb{R}^{F_i,K,2C}$. These tensors are produced by an MLP given the latent variable vector w_c^k:

$$\{\gamma_i[:,k,c], \beta_i[:,k,c]\} = \text{MLP}_{\text{STIN}}(w_c^k), \tag{7}$$

$$g_i^{'} = \text{STIN}(g_i, \gamma_i, \beta_i) = \begin{bmatrix} \gamma_i[1,k,c]\frac{g_i[1]-\mu(g_i[1])\mathbf{1}}{\sigma(g_i[1])} + \beta_i[1,k,c]\mathbf{1} \\ \vdots \\ \gamma_i[F_i,k,c]\frac{g_i[F_i]-\mu(g_i[F_i])\mathbf{1}}{\sigma(g_i[F_i])} + \beta_i[F_i,k,c]\mathbf{1} \end{bmatrix} \tag{8}$$

In Eq. (8), $\mathbf{1} \in \mathbb{R}^{H_j,W_j}$ is a matrix of ones, $\mu(\cdot)$, $\sigma(\cdot)$ calculate the sample-specific mean and std for each channel in the input feature map $g_i[j] \in \mathbb{R}^{H_i,W_i}$ (Ulyanov et al., 2017). Next, an STCA subblock modulates features maps in the channel dimension via a learnable attention tensor $R_i \in \mathbb{R}^{F_i,K,2C}$ (Li et al., 2023), produced by an MLP given w_c^k:

$$R_i[:,k,c] = \text{MLP}_{STCA}(w_c^k), \tag{9}$$

$$f_i = \text{STCA}(g_i', R_i) = \begin{bmatrix} g_i'[1] \odot R_i[1,k,c]\mathbf{1} \\ \vdots \\ g_i[F_i] \odot R_i[F_i,k,c]\mathbf{1} \end{bmatrix} \tag{10}$$

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

$$\hat{x}_{t_c} = \text{CB}_{L_S} \circ \text{PB}_{L_S-1} \circ \text{CB}_{L_S-1} \circ \cdots \circ \text{PB}_1 \circ \text{CB}_1(x_{s_c}, w_c^k), \tag{11}$$

where $\circ$ denotes functional composition, and $S_i := \text{PB}_i \circ \text{CB}_i$ for $i < L_s$ while $S_{L_s} := \text{CB}_{L_s}$ (i.e., the final decoder stage).

Discriminator (D): A local discriminator D^k with parameters θ_D^k is trained at site k based on a conditional patch-based architecture (Dar et al., 2019). Given the source image, D^k estimates the probability that x is an actual target image:

$$p_D = D^k(x, x^k_{s_c}), \tag{12}$$

where x is an actual or synthetic image of contrast t_c at site k.

3.2.2. *Partial network aggregation*

In pFLSynth, decentralized learning is performed for P communication rounds between the server and individual sites (Alg. 1). However, the conventional FedAvg algorithm aggregates the entire model across sites, increasing communication load and risk of information leakage (Elmas et al., 2022), while reducing sensitivity to site-specific features (Feng et al., 2021b). To address these limitations, here we perform PNA over later stages of the generator (Fig. 3). S is split at stage L_{cut} into two disjoint subsets $S_A = \{S_1, \dots, S_{L_{cut}}\}$ and $S_B = \{S_{L_{cut}+1}, \dots, S_{L_S}\}$ where the cut-point L_{cut} is selected from $\{1, \dots, L_S\}$. Earlier stages S_A with parameters $\theta^k_{S_A}, k \in \{1, \dots, K\}$ are kept locally at each site, whereas later stages S_B with parameters θ_{S_B} are shared, improving site specialization in source-image representations. Since we observed similar performance with local versus partially aggregated PBs, to lower communication costs and potential for information leakage, all PB$_i$ with parameters $\theta^k_{\text{PB}_i}$ are kept local and PNA is only exercised on CBs. These procedures elicit shared S_B and M, albeit unshared S_A and D across sites.

In the first round, the server randomly initializes global $\{S_B, M\}$ with parameters $\{\theta_{S_B}, \theta_M\}$. At the start of each round, the server broadcasts the global S_B and M to the sites:

$$\theta^k_{S_B} \leftarrow \theta_{S_B};\ \theta^k_M \leftarrow \theta_M; k = 1, 2, \dots, K \tag{13}$$

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

$$\theta^k_{S_A} \leftarrow \theta^k_{S_A};\ \theta^k_D \leftarrow \theta^k_D, \tag{14}$$

A local generator is then composed as:

$$G^k = \{(S_A^k \sqcup S_B^k), M^k\}. \tag{15}$$

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

$$\text{config}^k = \{(s_1, t_1), \dots, (s_C, t_C)\}, \tag{16}$$

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

$$\mathcal{L}^k_{syn}(\mathcal{D}^k, \theta^k) = \sum_{c=1}^{C} \mathbb{E}_{x^k_{s_c}, x^k_{t_c}} [-(D^k(x^k_{t_c}, x^k_{s_c}) - 1)^2 - \\ D^k(\hat{x}^k_{t_c}, x^k_{s_c})^2 + \lambda_{pix} \| x^k_{t_c} - \hat{x}^k_{t_c} \|_1]. \tag{17}$$

At the end of a round, each site sends its S_B and M to the server for aggregation:

$$\theta_{S_B} = \sum_{k=1}^{K} \frac{n^k}{n} \theta^k_{S_B};\ \theta_M = \sum_{k=1}^{K} \frac{n^k}{n} \theta^k_M. \tag{18}$$

In the final round, local generators are obtained by skipping aggregation (Cheng et al., 2021). During inference, each site uses its local generator to perform contrast translation:

$$\hat{x}^k_{t_c} = G^{k*}(x^k_{s_c}, v^k, u_c). \tag{19}$$

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

4. Experiments

4.1. *Datasets*

Experiments were conducted on four multi-contrast brain MRI datasets: IXI,[2] BRATS (Menze et al., 2015), MIDAS (Bullitt et al.,

[2] https://brain-development.org/ixi-dataset/

Algorithm 1: Training of pFLSynth

```
Data: $\{\mathcal{D}^1, \cdots, \mathcal{D}^K\}$ from $K$ sites
Input: $P$: number of communication rounds
$E$: number of local epochs
$G^1, \cdots, G^K$: local generators with params. $\theta_{G^1}, \cdots, \theta_{G^K}$
$D^1, \cdots, D^K$: local discriminators with $\theta_{D^1}, \cdots, \theta_{D^K}$
$S_B$, $M$: global generator components with $\theta_{S_B}, \theta_M$
$Opt()$: optimizer for parameter updates
$FedAvg()$: federated averaging
Output: $\theta^*_{G^k}$ personalized generators
Randomly initialize $\theta_{S_B}$, $\theta_M$ and $\theta_{D^1}, \cdots, \theta_{D^K}$
for $p = 1$ to $P$ do
    for $k = 1$ to $K$ do
        $\theta^k_{S_B} \leftarrow \theta_{S_B}$
    , $\theta^k_M \leftarrow \theta_M$                              // receive global
    for $e = 1$ to $E$ do
        Calculate $\nabla_{\theta^k_G} \mathcal{L}^k_{syn}(\mathcal{D}^k)$ based on Eq. (17)
        $\theta^k_G \leftarrow \theta^k_G - Opt(\nabla_{\theta^k_G} \mathcal{L}^k_{syn}(\mathcal{D}^k))$
        Calculate $\nabla_{\theta^k_D} \mathcal{L}^k_{syn}(\mathcal{D}^k)$ based on Eq. (17)
        $\theta^k_D \leftarrow \theta^k_D - Opt(\nabla_{\theta^k_D} \mathcal{L}^k_{syn}(\mathcal{D}^k))$
$\theta_{S_B,M} \leftarrow FedAvg(\theta^k_{S_B,M})$                       // aggregate
```

2005), and OASIS (LaMontagne et al., 2019). 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. Across the four datasets, the training set contained 2780, 2500, 3874, 2780 cross-sections, i.e., two-dimensional (2D) slices, per source-target configuration, respectively. Thus, given two source-target configurations per site, a total of 23,868 training cross-sections were used. Prior to modeling, each cross-section was normalized to a mean pixel intensity of 0.5, and the intensity range was clipped to [0 1]. Details about each dataset are provided below.

IXI Dataset: T_1-, T_2-, and Proton Density (PD)-weighted images from 53 subjects were analyzed with a (25,10,18) split. T_2- and PD-weighted images were registered onto T_1-weighted images via FSL using affine transformation based on mutual information (Jenkinson and Smith, 2001). For T_1, TE = 4.6 ms, TR = 9.8 ms, flip angle = 8°, spatial resolution = $0.94 \times 0.94 \times 1.2$ mm^3 were prescribed. For T_2, TR = 8178 ms, TE = 100 ms, flip angle = 90°, spatial resolution = $0.94 \times 0.94 \times 1.2$ mm^3 were prescribed. For PD, TR = 8178 ms, TE = 8 ms, flip angle = 90°, spatial resolution = $0.94 \times 0.94 \times 1.2$ mm^3 were prescribed.

BRATS Dataset: T_1-, T_2-, and Fluid Attenuation Inversion Recovery (FR)-weighted images from 55 subjects were analyzed with a (25,10,20) split. In BRATS, scans were acquired under various settings without a common scan protocol (Menze et al., 2015). As publicly shared, MR images were at $1 \times 1 \times 1$mm^3 resolution, skull-stripped, and co-registered to the same anatomical template.

MIDAS Dataset: T_1- and T_2-weighted images from 66 subjects were analyzed with a (48,5,13) split. MR images were co-registered to an anatomical template as publicly shared. For T_1, TR = 14 ms, TE = 7.7 ms, flip angle = 25°, spatial resolution = $1 \times 1 \times 1$ mm^3 were prescribed. For T_2, TR = 7730 ms, TE = 80 ms, flip angle = 90°, spatial resolution = $1 \times 1 \times 1$ mm^3 were prescribed.

OASIS Dataset: T_1-, T_2-, and FR-weighted images from 48 subjects were analyzed with a (22,9,17) split. T_2- and FR-weighted images were registered onto T_1-weighted images via FSL using affine transformation based on mutual information (Jenkinson and Smith, 2001). For T_1, TE = 4.0 ms, TR = 9.7 ms, flip angle = 10°, spatial resolution = $1 \times 1 \times 1$ mm^3 were prescribed. For T_2, TE = 86 ms, TR = 6150 ms, flip angle = 120°, spatial resolution = $1 \times 1 \times 1$ mm^3 were prescribed. For FR, TE = 91 ms, TR = 9000 ms, flip angle = 150°, spatial resolution = $1 \times 1 \times 1$ mm^3 were prescribed.

4.2. Competing methods

We demonstrated pFLSynth against centralized multi-site models, federated multi-site models, and single-site models. Centralized models were trained after forming an aggregate dataset across sites. Single-site models were trained using local data from each individual site. For each method, hyperparameter selection was performed via identical cross-validation procedures. All models employed 2D network architectures, and they shared generators across sites but used a separate local discriminator per site and per source-target configuration for improved performance.

4.2.1. Centralized multi-site models

pFLSynth$_{\text{cent}}$: A centralized version of pFLSynth was trained with the same architecture and loss function, albeit PNA was omitted. This model serves as a privacy-violating benchmark for pFLSynth.

pGAN$_{\text{cent}}$: A centralized version of the pGAN model with ResNet backbone was considered (Dar et al., 2019). The architecture and loss function were adopted from Dar et al. (2019).

pix2pix$_{\text{cent}}$: A centralized version of the pix2pix model with UNet backbone was considered (Isola et al., 2017). The architecture and loss function were adopted from Isola et al. (2017).

FedGAN$_{\text{cent}}$: A centralized version of a federated synthesis model was considered (Rasouli et al., 2020). The architecture and loss function were matched with pFLSynth, albeit the mapper and PBs were excluded.

4.2.2. Federated multi-site models

FedBN: A personalized model proposed for improved site-specialization in fMRI classification tasks was considered (Li et al., 2021). FedBN used the same architecture with pFLSynth, albeit it omitted the mapper and replaced PBs with site-specific batch normalization layers (Ioffe and Szegedy, 2015). For adversarial synthesis, a discriminator with matching loss function to pFLSynth was used.

FedMRI: A personalized model proposed for maintaining site-specialization in MRI reconstruction was considered (Feng et al., 2021b). FedMRI used a UNet backbone with a shared encoder and site-specific decoders (Feng et al., 2021b). For adversarial synthesis, a discriminator with matching loss function to pFLSynth was used.

FedGAN: A non-personalized model was implemented with matching loss function and architecture to pFLSynth, but without the mapper and PBs (Rasouli et al., 2020). FedGAN aggregated the entire generator.

FedMed: A non-personalized model for MRI synthesis was considered (Xie et al., 2022a). FedMed used a UNet generator that was entirely aggregated (Xie et al., 2022a). 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.

FedMM: A non-personalized task-unified MRI synthesis model was considered (Sharma and Hamarneh, 2020). FedMM was implemented with a U-Net generator and multiple input–output channels to cope with different source-target configurations. The entire generator was aggregated on the server.

FedCycle: A non-personalized task-unified model originally proposed for low-dose CT denoising was considered (Song and Ye, 2021). FedCycle used a UNet generator as in Song and Ye (2021), but only the forward mapping generator was retained with matching loss function to pFLSynth for paired synthesis. The switching mechanism was used to adapt the model to different source-target configurations.

4.2.3. Single-site models

pFLSynth$_{\text{sing}}$: A single-site version of pFLSynth was trained with the same architecture, albeit PNA was omitted.

FedGAN$_{\text{sing}}$: A single-site MRI synthesis model was considered (Dar et al., 2019), with the same architecture and loss function in pFLSynth, albeit the mapper and PBs excluded.

4.3. Architectural details

In pFLSynth, M was an MLP with $L_M = 6$ layers and sigmoid activation. S followed an encoder-bottleneck-decoder structure with $L_S = 15$ stages. The encoder had 3 stages ($e1 - e3$), each containing a cascade of a CB and a PB. Convolutional kernel sizes were 7, 3, 3 across stages. The bottleneck had 9 stages ($r1 - r9$), each containing a cascade of a residual CB with kernel size 3 (He et al., 2016) and a PB. The decoder had 3 stages ($d1 - d3$), the first two stages contained a cascade of a CB and a PB, and the last stage contained only a CB. Convolutional kernel sizes were 3, 3, 7. All CBs used ReLU activation except for the CB in the final decoder stage that used a tanh activation. The encoder mapped source images to feature maps in $\mathbb{R}^{256,64,64}$, the bottleneck retained dimensionality, and the decoder mapped feature maps back to $\mathbb{R}^{1,256,256}$. D had 5 convolutional layers with kernel size 4 and leaky ReLU activation, and an output layer with sigmoid activation. An FL setup with $K = 4$ different datasets was considered, so a site index of $v^k \in \mathbb{Z}_2^4$ was used. Given multiple source-task configurations, uniform random sample selection was utilized for learning different tasks. The datasets examined included T_1, T_2, PD and FR contrasts, so a source-target configuration index of $u_c \in \mathbb{Z}_2^8$ was formed. The mapper received these indices and produced a latent variable vector $w_c^k \in \mathbb{R}^{512}$. In each PB, the STIN subblock used a single-layer MLP with linear activation, and the STCA subblock used a two-layer MLP with sigmoid activation. Depending on the generator stage (i.e., i), MLP_{STIN} modules mapped w_c^k onto mean and std normalization parameters in $\mathbb{R}^{F_i}$, and MLP_{STCA} modules mapped w_c^k onto channel attention parameters in $\mathbb{R}^{F_i}$. The cut-point L_{cut} of the generator was selected as $r5$ based on validation performance.

4.4. Modeling procedures

For fair comparison, all models were trained using the discriminator described in Eq. (12) and the compound synthesis loss in Eq. (17). Hyperparameter selection, including weighting of the pixel-wise loss term, number of communication rounds, number of epochs, and learning rate was performed via cross-validation. A common set of hyperparameters that yielded near-optimal results across models and datasets were selected. Training was performed via Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. 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 $\lambda_{pix} = 100$. Models were trained and tested on cross-sections (i.e., 2D slices) within MRI volumes. Models were implemented using the PyTorch framework and executed on Nvidia RTX 3090 GPUs. Source codes of competing methods, scripts for data preparation and model evaluation, and a user guideline are available at: https://github.com/icon-lab/pFLSynth. In early phases of the study, we experimented with storing MR images in 32-bit versus 8-bit precision formats. Note that 32-bit precision was still used throughout the network models in both cases. As we did not observe notable differences between the two formats, we opted for the 8-bit format as commonly exercised with image-to-image translation models in computer vision.

Synthesis performance was evaluated via peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and Frechet inception distance (FID) metrics. PSNR and SSIM were measured on each cross-section, whereas FID yielded an aggregate measure across the test set. Segmentation performance was evaluated via Dice score and mean intersection over union (mIoU) metrics. Wilcoxon signed-rank tests were conducted to assess the significance of performance differences among competing methods ($p < 0.05$).

5. Results

5.1. Federated MRI synthesis performance

To demonstrate pFLSynth, FL experiments were conducted in a four-site setup based on IXI, BRATS, MIDAS, OASIS datasets taken as individual sites. First, we examined performance under implicit data heterogeneity in multi-site datasets by prescribing a common task configuration across sites ($T_1 \rightarrow T_2$ and $T_2 \rightarrow T_1$ in all sites). pFLSynth was compared against state-of-the-art federated baselines including personalized methods (FedBN, FedMRI), non-personalized methods (FedGAN, FedMed), and non-personalized task-unified methods (FedMM, FedCycle). Performance metrics listed in Table 2 indicate that pFLSynth outperforms federated baselines at each site ($p < 0.05$), except for BRATS where FedGAN yields lower FID and FedBN yields similar PSNR in $T_2 \rightarrow T_1$, and MIDAS where FedCycle yields higher PSNR and lower FID in $T_2 \rightarrow T_1$. On average across sites, pFLSynth achieves 1.2 dB higher PSNR, 2.9% higher SSIM, 9.7 lower FID than personalized baselines (FedBN, FedMRI), and 1.0 dB higher PSNR, 2.1% higher SSIM, 6.2 lower FID than non-personalized baselines (FedGAN, FedMed, FedMM, FedCycle). Note that pFLSynth generally offers competitive performance with its centrally-trained benchmark $\text{pFLSynth}_{\text{cent}}$, while outperforming centralized models based on other architectures.

Next, we examined performance under explicit data heterogeneity in multi-site datasets by prescribing a variable task configuration across sites ($T_1 \rightarrow T_2$ and $T_2 \rightarrow$PD in IXI, $T_1 \rightarrow T_2$ and FR$\rightarrow T_2$ in BRATS, $T_1 \rightarrow T_2$ and $T_2 \rightarrow T_1$ in MIDAS, $T_1 \rightarrow T_2$ and $T_2 \rightarrow$FR in OASIS, where FR denotes FLAIR). Performance metrics for competing methods are listed in Table 3. pFLSynth outperforms federated baselines at each site ($p < 0.05$), except for MIDAS where FedCycle yields similar PSNR in $T_2 \rightarrow T_1$and FedGAN performs similarly in $T_2 \rightarrow T_1$. On average across sites, pFLSynth achieves 1.5 dB higher PSNR, 3.6% higher SSIM, 9.1 lower FID than personalized baselines, and 2.7 dB higher PSNR, 6.8% higher SSIM, 31.8 lower FID than non-personalized baselines. Again, pFLSynth generally yields competitive performance with the centrally-trained benchmark $\text{pFLSynth}_{\text{cent}}$, while outperforming centralized models based on other architectures. Note that the benefits of pFLSynth over non-personalized baselines are relatively higher for the variable versus the common task configuration, indicating the elevated importance of the proposed personalization mechanisms in coping with explicit heterogeneity.

Representative images from pFLSynth and three top-contending federated baselines are displayed in Figs. 4 and 5 for the common and variable task configurations, respectively. Overall, federated baselines show a degree of noise-amplification or blurring artifacts, along with occasional inaccuracies in tissue structure. In contrast, pFLSynth yields lower artifacts and noise along with more accurate tissue depiction. These visual differences might be attributed to the technical differences between competing methods. Note that FedGAN in a non-personalized method that uses a global generator that can be rather insensitive to site-specific image features. Among personalized methods, FedMRI lacks normalization blocks to effectively modulate feature map statistics per site/task, and FedBN uses per-batch normalization parameters that can be suboptimally sensitive to sample-specific features. Thus, taken together, our results indicate that the personalization mechanisms embodied in pFLSynth based on sample-specific PBs and PNA can offer improved reliability against data heterogeneity in federated MRI synthesis.

5.2. Radiological evaluations

Radiological evaluations were conducted to examine the visual quality of images synthesized with federated models. An expert radiologist assessed the similarity of synthetic target images to ground-truth target images on a 5-point Likert scale. These assessments were performed for the common task configuration reported in Section 5.1, and

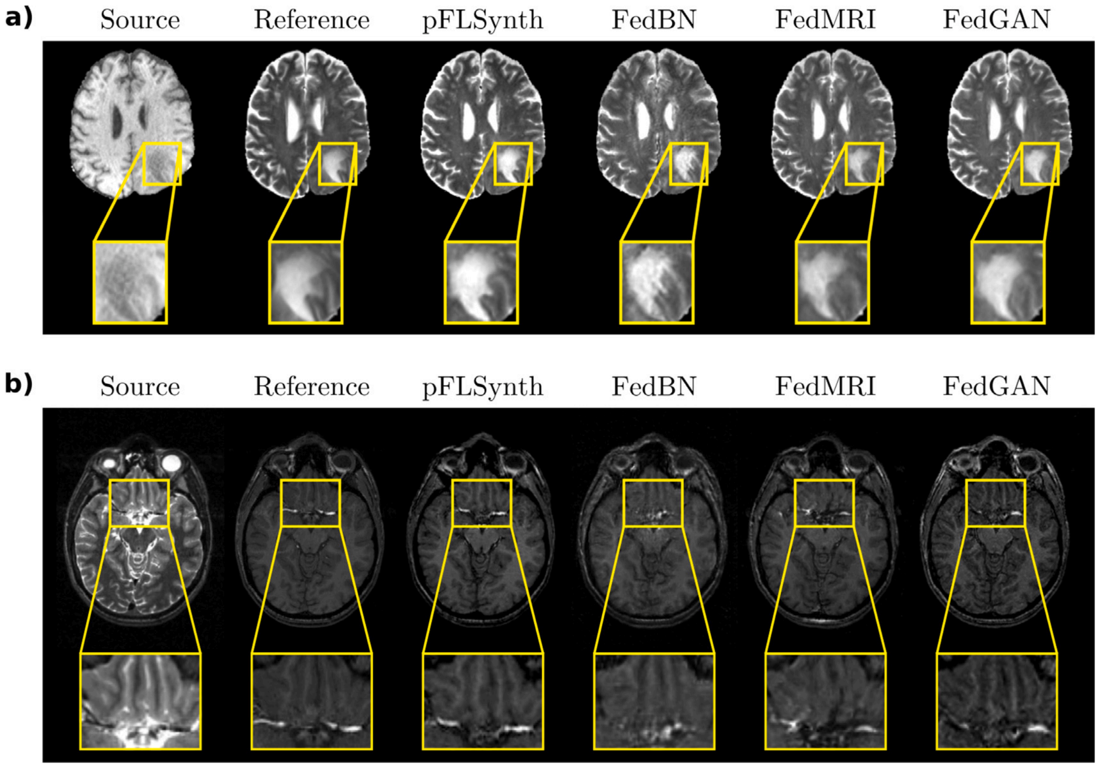


Fig. 4. Federated synthesis under a common task configuration across sites. Source, reference target, and synthetic target images from pFLSynth and three top-contending federated baselines are displayed (see Supp. Fig. 1 for all competing methods). Representative images for (a) $T_1 \rightarrow T_2$ in BRATS, (b) $T_2 \rightarrow T_1$ in MIDAS. Overall, pFLSynth synthesizes images with fewer artifacts and lower noise levels compared to competing federated methods.

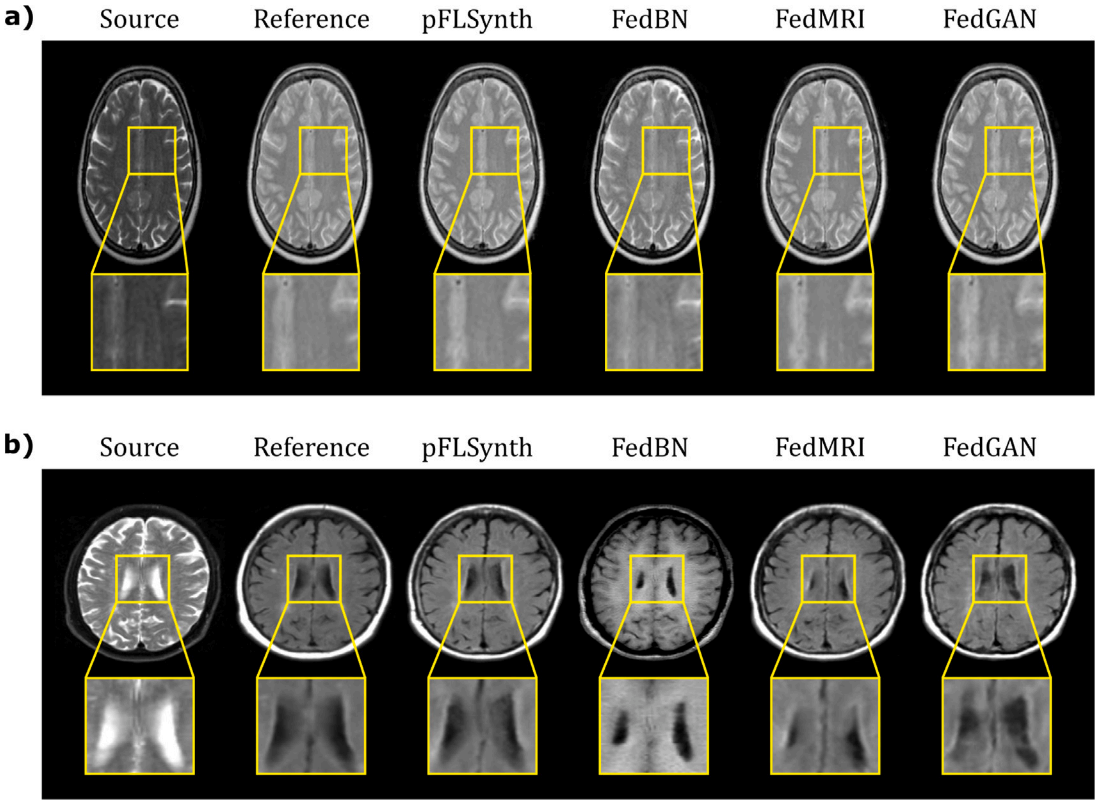


Fig. 5. Federated synthesis under a variable task configuration across sites. Source, reference target, and synthetic target images from pFLSynth and three top-contending federated baselines are displayed (see Supp. Fig. 2 for all competing methods). Representative images for (a) $T_2 \rightarrow PD$ in IXI and (b) $T_2 \rightarrow FR$ in OASIS. Overall, pFLSynth synthesizes images with fewer artifacts and lower noise levels compared to competing methods.

Table 2
Performance of federated models in a common task configuration of T_1→T_2 and T_2→T_1 in all sites. Centrally trained benchmarks pFLSynth$_{cent}$, pGAN$_{cent}$, and pix2pix$_{cent}$ are also reported. PSNR (dB), SSIM (%) are listed as mean±std across test subjects. The rightmost column lists the average metric across sites. Boldface indicates the top-performing federated model for each site, task, and metric.

			IXI		BRATS		MIDAS		OASIS		Average
			T_1→T_2	T_2→T_1	T_1→T_2	T_2→T_1	T_1→T_2	T_2→T_1	T_1→T_2	T_2→T_1	
Centralized models	pFLSynth$_{cent}$	PSNR ⇑	28.9 ± 1.1	27.9 ± 1.0	26.4 ± 0.8	25.0 ± 1.8	28.6 ± 0.5	26.5 ± 1.2	25.0 ± 0.5	21.7 ± 0.7	26.3
		SSIM ⇑	95.0 ± 1.4	94.6 ± 1.2	93.1 ± 0.9	93.3 ± 1.0	92.5 ± 0.9	87.2 ± 2.1	83.6 ± 2.2	77.8 ± 3.1	89.6
		FID ⇓	8.1	28.5	27.7	14.6	9.9	11.2	23.5	20.1	18.0
	pGAN$_{cent}$	PSNR ⇑	28.2 ± 1.2	27.8 ± 1.1	26.0 ± 0.8	24.4 ± 1.8	27.9 ± 0.6	26.1 ± 1.2	25.0 ± 0.5	21.3 ± 0.8	25.8
		SSIM ⇑	93.9 ± 1.4	94.4 ± 1.1	92.9 ± 1.0	92.4 ± 1.0	91.5 ± 1.0	86.9 ± 2.2	83.2 ± 2.4	76.9 ± 2.0	89.0
		FID ⇓	8.6	28.5	28.5	14.6	10.4	11.8	24.8	24.1	18.9
	pix2pix$_{cent}$	PSNR ⇑	26.3 ± 0.9	27.0 ± 1.0	25.8 ± 0.7	24.9 ± 1.8	26.9 ± 0.5	25.1 ± 1.0	23.9 ± 0.6	21.3 ± 0.7	25.2
		SSIM ⇑	89.7 ± 1.6	91.7 ± 1.3	90.8 ± 1.0	91.2 ± 1.1	85.2 ± 1.3	84.3 ± 2.2	81.1 ± 3.6	74.0 ± 3.4	86.0
		FID ⇓	26.8	35.4	40.0	20.7	21.3	21.8	32.8	36.0	29.4
Federated models	pFLSynth	PSNR ⇑	**28.6 ± 1.3**	**28.0 ± 1.1**	**26.3 ± 0.8**	**24.7 ± 1.7**	**28.4 ± 0.6**	26.2 ± 1.3	**25.0 ± 0.6**	**21.4 ± 0.8**	**26.1**
		SSIM ⇑	**94.5 ± 1.3**	**94.9 ± 1.2**	**93.0 ± 1.0**	**93.0 ± 1.0**	**92.0 ± 0.8**	**86.5 ± 2.4**	**83.7 ± 2.4**	**77.1 ± 3.0**	**89.3**
		FID ⇓	**8.5**	**26.5**	26.5	14.3	**10.0**	11.8	**32.0**	**21.0**	**18.8**
	FedBN	PSNR ⇑	27.1 ± 0.9	26.0 ± 0.8	25.5 ± 1.1	24.6 ± 0.9	27.7 ± 0.5	24.8 ± 1.0	24.6 ± 0.5	20.5 ± 0.8	25.1
		SSIM ⇑	91.6 ± 1.8	90.6 ± 1.5	91.9 ± 1.4	92.1 ± 1.3	89.1 ± 2.3	85.3 ± 3.0	81.2 ± 3.2	73.2 ± 3.1	86.9
		FID ⇓	13.4	30.0	26.4	15.1	12.0	11.7	42.2	36.4	23.4
	FedMRI	PSNR ⇑	27.0 ± 1.0	26.7 ± 1.0	25.4 ± 0.9	24.2 ± 1.5	27.1 ± 0.5	25.4 ± 1.0	22.6 ± 0.5	19.4 ± 0.6	24.7
		SSIM ⇑	92.9 ± 1.4	93.5 ± 1.3	92.2 ± 1.1	91.6 ± 1.0	90.2 ± 1.2	85.3 ± 1.9	73.4 ± 2.1	68.2 ± 2.6	85.9
		FID ⇓	13.6	35.2	35.3	16.6	13.2	18.5	80.6	55.7	33.6
	FedGAN	PSNR ⇑	26.9 ± 1.0	26.5 ± 0.8	25.6 ± 0.8	24.3 ± 1.9	27.4 ± 0.5	25.5 ± 1.0	24.2 ± 0.5	20.6 ± 0.7	25.1
		SSIM ⇑	92.3 ± 1.5	92.9 ± 1.4	91.4 ± 1.0	92.3 ± 1.0	89.2 ± 1.0	84.7 ± 1.9	81.2 ± 1.9	75.3 ± 2.9	87.4
		FID ⇓	14.5	32.1	**23.1**	**11.9**	10.6	12.6	43.1	45.8	24.2
	FedMed	PSNR ⇑	26.6 ± 1.1	26.5 ± 1.1	25.3 ± 0.8	24.2 ± 1.6	27.3 ± 0.5	25.5 ± 1.1	23.6 ± 0.6	20.5 ± 0.6	24.9
		SSIM ⇑	92.2 ± 1.7	93.1 ± 1.3	92.2 ± 0.9	91.9 ± 0.9	90.6 ± 1.0	85.8 ± 1.8	76.6 ± 2.5	74.1 ± 2.9	87.1
		FID ⇓	14.3	38.4	29.5	14.8	10.3	12.1	56.4	31.2	25.9
	FedMM	PSNR ⇑	26.4 ± 1.1	26.6 ± 1.0	25.4 ± 0.7	24.3 ± 1.7	27.5 ± 0.5	25.6 ± 1.1	23.6 ± 0.5	20.5 ± 0.6	25.0
		SSIM ⇑	91.5 ± 1.8	92.7 ± 1.3	92.3 ± 1.0	91.2 ± 1.0	90.7 ± 0.9	86.1 ± 1.9	78.3 ± 2.6	73.7 ± 3.0	87.1
		FID ⇓	20.4	39.2	32.3	13.7	10.5	12.3	45.1	30.3	25.5
	FedCycle	PSNR ⇑	26.9 ± 1.0	26.5 ± 1.0	25.5 ± 0.8	24.4 ± 1.7	27.5 ± 0.5	**27.7 ± 1.2**	23.3 ± 0.7	20.6 ± 0.8	25.3
		SSIM ⇑	92.7 ± 1.3	93.3 ± 1.3	92.5 ± 1.0	92.2 ± 1.0	91.0 ± 1.0	86.3 ± 2.1	76.6 ± 2.5	74.6 ± 3.0	87.4
		FID ⇓	13.4	33.9	29.4	13.7	10.1	**11.6**	51.6	33.3	24.6

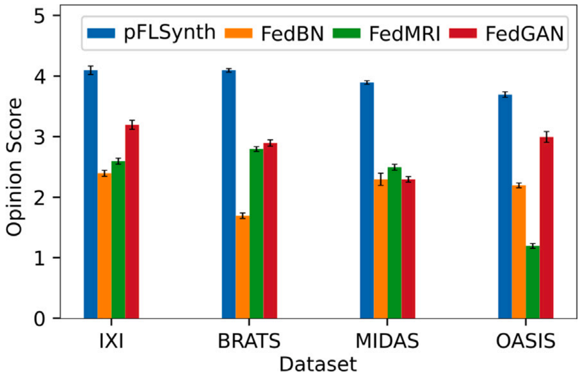


Fig. 6. Radiological opinion scores for pFLSynth, FedBN, FedMRI, and FedGAN for the common task configuration (T_1→T_2 and T_2→T_1 at all sites). For each method, mean±se of opinion scores across tasks are shown.

on a set of 10 randomly selected test subjects per site. Radiological opinion scores of pFLSynth and three top-contending federated baselines are displayed in Fig. 6. We find that pFLSynth outperforms all federated baselines at each site ($p < 0.05$). On average across sites, pFLSynth achieves an opinion score of 4.0, whereas FedBN, FedMRI, and FedGAN yield lower opinion scores of 2.2, 2.3, and 2.9, respectively. The relatively low scores of federated baselines were driven by the presence of several prominent image artifacts. In particular, federated baselines showed blurring artifacts near the cerebrospinal fluid (CSF)-parenchyma interface, consistently across datasets. There were noise-amplification artifacts with fine granular appearance in MIDAS and coarse granular appearance in OASIS. Structural inaccuracies due to image-degradation artifacts were apparent paramedially along the anterior–posterior direction in IXI and BRATS, transversally along the right-left direction in MIDAS, and diffusely across the brain in OASIS. In comparison to baselines, pFLSynth showed lower artifacts, higher gray-white matter differentiation and better structural delineation of brain regions including the basal ganglia. These visual differences can be important factors for interpretation of brain anatomy as well as lesion detection and characterization during radiological evaluations. As such, these results suggest that quantitative improvements in synthetic images generated by pFLSynth are accompanied by clinically-relevant improvements in visual quality.

5.3. Segmentation based on synthetic images

Next, we examined the utility of synthetic images generated by federated models in a downstream segmentation task. This analysis was conducted on the BRATS dataset that includes ground-truth segmentation masks for brain tumors (Menze et al., 2015). A UNet-based model[3] pre-trained for brain tumor segmentation on The Cancer Genome Atlas (TCGA) lower-grade glioma collection was employed (Buda et al., 2019). The segmentation model expected T_1-, T_2- and FLAIR-weighted images as input. Assuming a scenario where T_2-weighted images were not acquired, synthetic T_2-weighted images were generated by executing the T_1→T_2 task, and these synthetic images were then provided to the segmentation model along with actual T_1- and FLAIR-weighted images. Performance metrics for tumor segmentation based on synthetic images imputed by competing methods are listed in Table 4. We find that pFLSynth-generated images yield more accurate segmentations

[3] https://github.com/Th3NiKo/Brain-MRI-segmentation-unet

Table 3
Performance of federated models in a variable task configuration across sites. Centrally trained benchmarks are also reported. FR denotes FLAIR.

			IXI		BRATS		MIDAS		OASIS		Average
			$T_1 \rightarrow T_2$	$T_2 \rightarrow$ PD	$T_1 \rightarrow T_2$	FR$\rightarrow T_2$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	$T_1 \rightarrow T_2$	$T_2 \rightarrow$ FR	
Centralized models	pFLSynth$_{cent}$	PSNR ⇑	28.4 ± 1.1	31.5 ± 0.9	26.0 ± 0.9	24.9 ± 1.1	26.2 ± 0.5	25.6 ± 1.0	23.9 ± 1.6	22.0 ± 2.0	26.1
		SSIM ⇑	94.6 ± 1.4	97.6 ± 0.4	93.4 ± 1.1	90.9 ± 1.6	89.8 ± 1.2	85.1 ± 2.2	82.8 ± 3.3	80.2 ± 5.8	89.3
		FID ⇓	9.4	23.1	22.6	24.6	9.2	11.9	30.2	23.7	19.3
	pGAN$_{cent}$	PSNR ⇑	25.2 ± 1.3	26.8 ± 1.8	24.5 ± 1.0	21.8 ± 1.4	22.7 ± 1.0	20.8 ± 1.0	24.3 ± 1.2	21.9 ± 1.4	23.5
		SSIM ⇑	89.6 ± 2.9	86.6 ± 6.4	90.4 ± 1.1	81.7 ± 5.1	70.2 ± 6.7	67.1 ± 2.7	81.6 ± 3.6	79.8 ± 3.6	80.9
		FID ⇓	29.1	59.1	45.3	85.1	48.0	11.5	39.1	26.4	43.0
	pix2pix$_{cent}$	PSNR ⇑	26.5 ± 0.9	29.2 ± 1.1	24.5 ± 0.9	21.4 ± 1.0	25.4 ± 0.7	24.1 ± 0.8	22.4 ± 1.6	20.7 ± 1.4	24.3
		SSIM ⇑	89.0 ± 1.8	94.1 ± 0.7	86.9 ± 1.7	80.5 ± 2.9	80.8 ± 2.3	69.7 ± 2.2	69.1 ± 3.9	66.2 ± 4.8	79.5
		FID ⇓	19.3	29.2	51.9	83.9	21.6	27.0	70.5	62.5	45.7
Federated models	pFLSynth	PSNR ⇑	**28.3 ± 1.2**	**31.5 ± 1.0**	**26.0 ± 0.9**	**24.3 ± 1.0**	**26.7 ± 0.6**	25.4 ± 1.1	**23.7 ± 1.6**	**21.8 ± 1.9**	**26.0**
		SSIM ⇑	**94.2 ± 1.4**	**97.4 ± 0.5**	**93.5 ± 1.1**	**90.5 ± 1.6**	**90.7 ± 1.1**	85.3 ± 2.0	**82.2 ± 3.1**	**79.2 ± 4.7**	**89.1**
		FID ⇓	**9.1**	**21.6**	**22.9**	**30.3**	**9.7**	**12.0**	**36.8**	**30.1**	**21.6**
	FedBN	PSNR ⇑	27.2 ± 0.5	30.6 ± 0.9	25.1 ± 0.6	23.1 ± 0.6	25.5 ± 0.5	23.9 ± 0.9	22.3 ± 1.0	19.9 ± 1.1	24.7
		SSIM ⇑	92.5 ± 2.9	96.4 ± 0.9	92.4 ± 2.6	88.4 ± 2.6	81.4 ± 1.8	79.5 ± 2.8	78.8 ± 4.9	76.4 ± 5.0	85.7
		FID ⇓	12.4	27.3	24.3	40.3	15.9	20.7	41.0	37.0	27.4
	FedMRI	PSNR ⇑	27.0 ± 1.1	30.7 ± 0.9	23.7 ± 1.1	22.7 ± 0.9	24.9 ± 0.5	23.1 ± 0.9	21.8 ± 1.3	20.0 ± 1.5	24.2
		SSIM ⇑	92.7 ± 1.5	96.6 ± 0.5	89.8 ± 1.7	87.8 ± 1.6	86.7 ± 1.2	79.0 ± 2.2	77.7 ± 2.7	72.3 ± 5.0	85.3
		FID ⇓	12.6	26.0	43.4	56.6	21.6	33.8	40.4	38.0	34.1
	FedGAN	PSNR ⇑	26.8 ± 1.1	29.7 ± 0.9	23.9 ± 1.1	20.1 ± 1.5	20.7 ± 0.9	25.4 ± 1.0	22.0 ± 1.6	19.9 ± 1.8	23.6
		SSIM ⇑	92.2 ± 1.7	96.2 ± 0.6	89.9 ± 1.5	81.5 ± 2.7	76.4 ± 3.2	**85.5 ± 2.1**	78.2 ± 3.3	75.2 ± 5.5	84.4
		FID ⇓	14.4	30.0	42.4	107.0	117.2	14.3	47.7	40.1	51.6
	FedMed	PSNR ⇑	26.2 ± 0.9	28.7 ± 0.8	22.8 ± 1.1	20.1 ± 1.0	25.5 ± 0.7	24.0 ± 1.4	21.4 ± 1.2	19.6 ± 1.8	23.5
		SSIM ⇑	91.5 ± 1.6	95.1 ± 0.7	88.2 ± 1.6	82.3 ± 2.0	88.9 ± 1.2	80.8 ± 2.5	75.7 ± 3.3	71.2 ± 5.3	84.2
		FID ⇓	18.8	29.4	59.9	121.4	19.9	17.3	53.9	39.9	45.1
	FedMM	PSNR ⇑	27.1 ± 1.0	30.3 ± 0.8	24.2 ± 0.8	22.9 ± 0.9	17.5 ± 0.8	24.0 ± 1.4	21.7 ± 1.3	19.2 ± 1.3	23.4
		SSIM ⇑	92.7 ± 1.6	96.5 ± 0.5	90.5 ± 1.1	87.9 ± 1.8	62.5 ± 3.2	80.8 ± 2.5	72.6 ± 4.2	68.8 ± 6.2	81.5
		FID ⇓	13.4	24.8	36.4	46.6	169.6	26.9	45.4	39.7	50.4
	FedCycle	PSNR ⇑	26.8 ± 1.0	28.7 ± 0.8	23.5 ± 0.9	18.9 ± 1.0	17.3 ± 0.9	**25.5 ± 1.0**	21.6 ± 1.3	19.7 ± 1.4	22.8
		SSIM ⇑	92.5 ± 1.5	95.6 ± 0.7	89.3 ± 1.0	80.1 ± 2.7	62.0 ± 3.2	84.0 ± 2.2	65.7 ± 5.1	64.0 ± 8.4	79.2
		FID ⇓	16.5	31.0	56.1	139.1	176.8	14.0	49.8	47.9	66.4

Table 4
Segmentation performance based on synthetic images imputed by competing methods on BRATS. Synthesis models trained for the common task configuration were used to synthesize T_2-weighted images by executing $T_1 \rightarrow T_2$. Synthetic T_2-weighted and actual T_1-, FLAIR-weighted images were provided to a pre-trained segmentation model to predict tumor regions. Dice score (%) and mIoU (%) are reported as mean±se across test subjects.

	pFLSynth	FedBN	FedMRI	FedGAN	FedMed	FedMM	FedCycle
Dice	**77.2** ± **5.2**	63.0 ± 7.1	63.6 ± 6.8	64.3 ± 6.8	64.2 ± 6.6	61.4 ± 7.2	65.6 ± 6.5
mIoU	**62.9** ± **7.0**	51.2 ± 7.1	52.4 ± 6.9	53.0 ± 7.0	52.4 ± 6.9	50.3 ± 7.5	50.7 ± 6.7

Table 5
Computational complexity of competing methods. N_{model}: number of local model parameters (millions), N_{comm}: number of transferred model parameters (millions), T_{train}: training time per cross-section (msec), T_{inf}: inference time per cross-section (msec), Memory: VRAM use (GB).

	pFLSynth	FedBN	FedMRI	FedGAN	FedMed	FedMM	FedCycle
N_{model}	18.75	14.27	10.51	14.27	10.51	10.52	12.75
N_{comm}	6.52	11.49	4.71	11.51	7.77	7.78	9.99
T_{train}	114.93	105.93	104.25	105.94	104.26	104.71	106.94
T_{inf}	14.10	12.48	5.13	12.48	5.13	5.53	7.44
Memory	1.73	1.67	1.63	1.67	1.63	1.64	1.65

than those generated by federated baselines. These results underscore the potential of pFLSynth in improving performance in downstream tasks following image formation.

5.4. Computational complexity

Practical concerns for decentralized model training include the model size that is transferred between individual sites and the server, and the local training times. Table 5 lists for competing methods the number of parameters for the overall model (N_{model}) and for the aggregated components within the model communicated with the server (N_{comm}), along with the training times per cross-section (T_{train}). Note that all methods retain local discriminators. FedGAN, FedMed, FedMM, and FedCycle transfer the generator entirely and FedBN transfers the generator apart from compact BN layers, resulting in a substantial portion of the overall model to be communicated. In contrast, FedMRI and pFLSynth adopt partial network aggregation to significantly lower communication load. All methods have comparable training times, with pFLSynth yielding modestly longer training. Meanwhile, practical concerns for inference include the inference times and memory use of federated synthesis models. Table 5 lists for competing methods the inference times (T_{inf}) and memory use (Memory) per cross-section. Given its relatively higher number of parameters, pFLSynth yields moderately longer albeit reasonable inference times, and it has comparable memory use to federated baselines.

5.5. Federated versus single-site models

Federated multi-site models trained on large, diverse datasets promise improved performance over single-site models trained on

Table 6
Performance for pFLSynth and ablated variants. A variant that ablated STIN subblocks (w/o STIN), a variant that ablated STCA subblocks (w/o STCA[a]), a variant based on a two-layer MLP$_{STIN}$ while ablating STCA subblocks (w/o STCA[b]), a variant that ablated PNA (w/o PNA), and a variant that ablated the mapper (w/o Mapper) were examined. Boldface indicates the top-performing model for each site, task, and metric.

		IXI		BRATS		MIDAS		OASIS		Average
		$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	
pFLSynth	PSNR ⇑	**28.6 ± 1.3**	**28.0 ± 1.1**	**26.3 ± 0.8**	**24.7 ± 1.7**	**28.4 ± 0.6**	**26.2 ± 1.3**	**25.0 ± 0.6**	**21.4 ± 0.8**	**26.1**
	SSIM ⇑	**94.5 ± 1.3**	**94.9 ± 1.2**	**93.0 ± 1.0**	**93.0 ± 1.0**	**92.0 ± 0.8**	**86.5 ± 2.4**	**83.7 ± 2.4**	**77.1 ± 3.0**	**89.3**
	FID ⇓	**8.5**	**26.5**	**26.5**	14.3	**10.0**	11.8	32.0	**21.0**	**18.8**
w/o STIN	PSNR ⇑	27.8 ± 1.1	27.5 ± 1.0	25.9 ± 0.8	24.3 ± 1.7	27.7 ± 0.5	26.0 ± 1.2	24.6 ± 0.6	21.0 ± 0.7	25.6
	SSIM ⇑	93.1 ± 1.5	91.7 ± 1.3	91.7 ± 1.3	92.1 ± 1.3	89.8 ± 1.6	84.7 ± 2.9	80.5 ± 3.5	75.3 ± 3.4	87.4
	FID ⇓	9.5	29.4	27.0	14.3	11.1	12.4	38.5	18.1	20.0
w/o STCA[a]	PSNR ⇑	27.8 ± 1.1	27.2 ± 1.0	25.9 ± 0.8	24.3 ± 1.7	27.6 ± 0.5	25.9 ± 1.1	24.5 ± 0.8	21.1 ± 0.7	25.5
	SSIM ⇑	92.9 ± 1.7	93.2 ± 1.5	91.6 ± 1.3	91.4 ± 1.3	89.6 ± 1.6	85.6 ± 2.8	81.7 ± 3.8	75.6 ± 3.4	87.7
	FID ⇓	10.0	31.2	29.9	13.1	10.3	12.5	33.3	21.4	20.1
w/o STCA[b]	PSNR ⇑	28.4 ± 1.2	**28.0 ± 1.1**	25.8 ± 1.3	23.9 ± 0.9	28.2 ± 0.6	26.0 ± 1.2	24.5 ± 0.7	20.8 ± 0.8	25.7
	SSIM ⇑	92.3 ± 1.7	93.1 ± 1.4	92.7 ± 1.5	92.7 ± 1.6	90.9 ± 1.2	85.8 ± 2.9	82.1 ± 3.3	76.2 ± 3.0	88.2
	FID ⇓	8.2	27.3	28.6	23.6	**10.0**	11.7	33.3	21.8	20.6
w/o PNA	PSNR ⇑	27.8 ± 1.1	26.9 ± 0.9	25.9 ± 0.9	24.5 ± 1.8	27.6 ± 0.5	25.8 ± 1.2	24.7 ± 0.7	21.2 ± 0.7	25.6
	SSIM ⇑	93.6 ± 1.9	93.3 ± 1.5	92.8 ± 1.3	92.3 ± 1.3	91.6 ± 1.6	86.3 ± 2.0	82.3 ± 3.2	76.6 ± 3.5	88.6
	FID ⇓	10.9	30.9	27.9	**12.2**	10.3	**11.5**	**27.5**	21.9	19.1
w/o Mapper	PSNR ⇑	27.5 ± 1.1	27.2 ± 0.9	12.4 ± 0.3	15.1 ± 0.8	27.4 ± 0.5	25.7 ± 1.0	24.5 ± 0.7	21.0 ± 0.7	22.6
	SSIM ⇑	92.7 ± 1.7	93.3 ± 1.4	74.3 ± 1.6	83.2 ± 1.2	90.6 ± 1.8	85.7 ± 3.1	80.7 ± 3.5	76.0 ± 3.3	84.6
	FID ⇓	10.1	30.7	103.9	90.4	11.5	12.0	37.0	19.9	39.4

compact local datasets. However, the extent of potential improvements depend inherently on the ability of FL methods in maintaining sensitivity to site-specific features while seeking cross-site generalization. To systematically assess the benefits of federated MRI synthesis, we evaluated synthesis performance as a function of the size of training sets. For this analysis, pFLSynth and FedGAN models were compared against their centralized (pFLSynth$_{cent}$, FedGAN$_{cent}$) and single-site versions (pFLSynth$_{sing}$, FedGAN$_{sing}$). Federated models were decentrally trained on multi-site datasets, centralized models were trained on aggregated multi-site datasets, and single-site models were trained separately on the local datasets for each site. Note that the centralized benchmarks denote an upper performance bound for underlying federated models. Experiments were conducted for the common task configuration. The relative size of the training set was varied in $R_{train} = \{0.125, 0.25, 0.5, 1\}$ with respect to the size of the original training sets by randomly selecting a subset of subjects. All models were trained on the same selected set of subjects for a given R_{train}, and they were tested on the same test subjects.

Fig. 7 displays average synthesis performance across sites as a function of R_{train}. Naturally, the benefits of federated learning become more prominent for both methods as the training set size for individual sites is reduced. However, while FedGAN yields higher PSNR than FedGAN$_{sing}$ at R_{train} = 0.125, it performs relatively poorly for remaining R_{train} values as the training set size grows (p < 0.05). In general, FedGAN performs poorly when compared against the centralized benchmark FedGAN$_{cent}$. These results suggest that FedGAN shows suboptimal sensitivity to site-specific features that prevent it to effectively benefit from learning on multi-site datasets. In contrast, pFLSynth outperforms pFLSynth$_{sing}$ across all R_{train} values ($p < 0.05$). Meanwhile, pFLSynth yields on par performance with the centralized benchmark pFLSynth$_{cent}$ across R_{train} values. These findings indicate that pFLSynth can effectively leverage the diverse information in multi-site datasets to maintain high generalization and specialization.

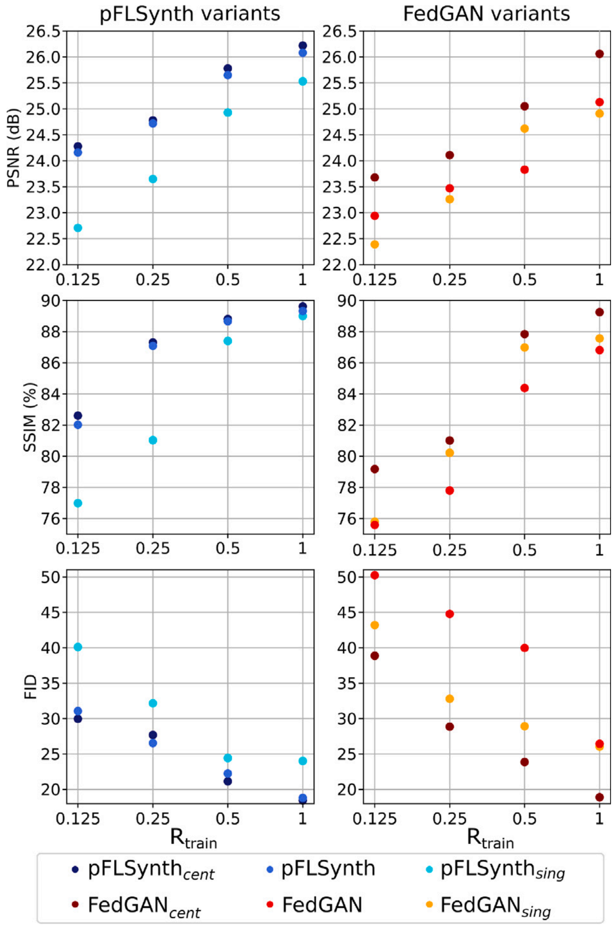


Fig. 7. Average synthesis performance across sites as a function of relative training set size, R_{train}. PSNR (top row), SSIM (middle row), and FID (bottom row) are plotted. Federated, centralized and single-site variants of pFLSynth are shown (left column), along with federated, centralized and single-site variants of FedGAN (right column).

5.6. Ablation studies

Ablation studies were conducted to assess the contributions of major design elements in pFLSynth to synthesis performance. First, pFLSynth was compared against variants where STIN subblocks were ablated (w/o STIN), STCA subblocks were ablated (w/o STCA[a]), STCA subblocks were removed and a two-layer MLP$_{STIN}$ was used (w/o STCA[b]), PNA was ablated by using full network aggregation (w/o PNA), and the mapper was ablated by inputting site and source-target configuration indices directly to PBs (w/o Mapper). Performance metrics for the common task configuration are listed in Table 6. pFLSynth consistently

Table 7
Performance for pFLSynth and variant models ablated of site index (v^k) and source-target configuration index (u_c). Boldface indicates the top-performing model for each site, task, and metric. FR denotes FLAIR.

		IXI		BRATS		MIDAS		OASIS		Average
		$T_1 \rightarrow T_2$	$T_2 \rightarrow PD$	$T_1 \rightarrow T_2$	$FR \rightarrow T_2$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	$T_1 \rightarrow T_2$	$T_2 \rightarrow FR$	
pFLSynth	PSNR ⇑	**28.3 ± 1.2**	**31.5 ± 1.0**	**26.0 ± 0.9**	**24.3 ± 1.0**	**26.7 ± 0.6**	**25.4 ± 1.1**	**23.7 ± 1.6**	**21.8 ± 1.9**	**26.0**
	SSIM ⇑	**94.2 ± 1.4**	**97.4 ± 0.5**	**93.5 ± 1.1**	**90.5 ± 1.6**	**90.7 ± 1.1**	**85.3 ± 2.0**	**82.2 ± 3.1**	**79.2 ± 4.7**	**89.1**
	FID ⇓	**9.1**	**21.6**	**22.9**	**30.3**	**9.7**	**12.0**	**36.8**	30.1	**21.6**
w/o v^k	PSNR ⇑	27.7 ± 1.1	30.8 ± 1.2	25.4 ± 1.1	18.6 ± 0.8	25.3 ± 0.5	24.0 ± 1.1	23.1 ± 1.5	20.1 ± 1.6	24.4
	SSIM ⇑	93.7 ± 1.4	96.8 ± 0.4	91.7 ± 1.2	80.8 ± 2.3	89.4 ± 1.5	82.2 ± 1.2	75.2 ± 4.8	71.0 ± 6.6	85.1
	FID ⇓	14.8	22.3	36.2	147.9	17.7	24.7	49.6	37.7	43.9
w/o u_c	PSNR ⇑	27.5 ± 1.2	30.5 ± 1.91	24.8 ± 1.3	7.2 ± 9.4	25.7 ± 0.6	25.3 ± 0.8	22.0 ± 1.5	20.7 ± 2.2	23.0
	SSIM ⇑	93.9 ± 1.1	97.2 ± 0.3	91.3 ± 1.0	29.7 ± 2.5	88.7 ± 0.9	84.8 ± 2.1	80.4 ± 3.0	74.9 ± 5.3	80.1
	FID ⇓	10.1	22.4	23.1	163.1	10.5	12.9	47.1	**29.9**	39.9

Table 8
Performance for pFLSynth variants based on different cut-points (L_{cut}) within the bottleneck of the generator. $L_{cut} = r5$ is the cut-point selected based on validation performance, and used in the main experiments.

		IXI		BRATS		MIDAS		OASIS		Average
		$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	$T_1 \rightarrow T_2$	$T_2 \rightarrow T_1$	
$r3$	PSNR ⇑	28.0 ± 1.0	27.7 ± 0.9	26.2 ± 0.9	**25.1 ± 1.8**	27.7 ± 0.5	26.1 ± 1.2	**25.6 ± 0.6**	**21.9 ± 0.7**	26.0
	SSIM ⇑	93.1 ± 1.4	93.9 ± 1.2	92.5 ± 1.2	**93.7 ± 1.1**	90.2 ± 1.7	**87.2 ± 2.5**	**84.6 ± 3.3**	77.1 ± 2.7	89.0
	FID ⇓	10.5	30.1	25.0	**14.1**	11.9	13.4	**30.7**	**20.8**	19.6
$r4$	PSNR ⇑	27.9 ± 1.0	27.8 ± 0.9	26.3 ± 1.0	24.8 ± 2.0	28.1 ± 0.7	26.0 ± 1.2	24.9 ± 0.7	21.4 ± 0.9	25.9
	SSIM ⇑	94.0 ± 1.4	94.0 ± 1.2	92.5 ± 1.1	93.5 ± 1.1	91.4 ± 1.5	85.8 ± 2.6	82.3 ± 3.2	76.5 ± 2.9	88.8
	FID ⇓	10.1	31.0	25.3	**14.1**	11.8	13.1	33.1	22.1	20.1
$r5$ (selected)	PSNR ⇑	**28.6 ± 1.3**	**28.0 ± 1.1**	26.3 ± 0.8	24.7 ± 1.7	**28.4 ± 0.6**	**26.2 ± 1.3**	25.0 ± 0.6	21.4 ± 0.8	**26.1**
	SSIM ⇑	**94.5 ± 1.3**	**94.9 ± 1.2**	93.0 ± 1.0	93.0 ± 1.0	**92.0 ± 0.8**	86.5 ± 2.4	83.7 ± 2.4	77.1 ± 3.0	**89.3**
	FID ⇓	**8.5**	**26.5**	26.5	14.3	**10.0**	**11.8**	32.0	21.0	**18.8**
$r6$	PSNR ⇑	28.1 ± 1.1	27.6 ± 1.3	26.4 ± 1.0	24.6 ± 1.6	28.1 ± 0.5	26.2 ± 1.2	24.7 ± 0.8	21.7 ± 0.6	25.9
	SSIM ⇑	93.8 ± 1.3	93.5 ± 1.3	**93.3 ± 1.2**	92.3 ± 1.1	91.7 ± 1.4	86.1 ± 2.7	83.4 ± 2.8	**77.4 ± 3.0**	88.9
	FID ⇓	9.1	29.5	**24.2**	18.5	10.4	12.3	31.2	21.4	19.6
$r7$	PSNR ⇑	28.3 ± 1.1	27.9 ± 0.9	**26.5 ± 1.1**	24.5 ± 2.0	28.4 ± 0.8	26.2 ± 1.2	24.2 ± 0.8	21.4 ± 1.0	25.9
	SSIM ⇑	93.5 ± 1.4	93.2 ± 1.2	92.2 ± 1.1	92.0 ± 1.1	91.7 ± 1.2	86.8 ± 2.7	82.2 ± 3.2	77.0 ± 3.0	88.6
	FID ⇓	11.8	31.5	27.2	19.0	11.0	12.2	33.3	21.6	21.0

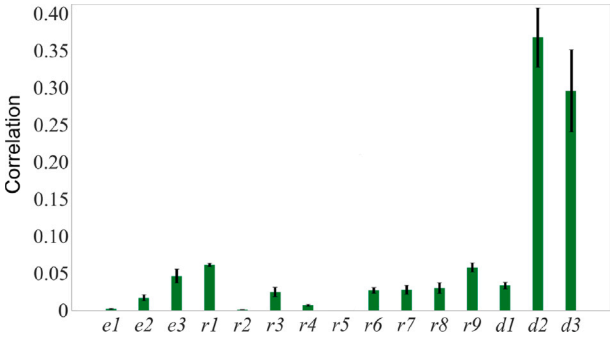


Fig. 8. Similarity of weight vectors across generator stages, shown as mean±std of Spearman's correlation coefficient across sites. Single-site pFLSynth$_{sing}$ models were trained separately on IXI, BRATS, MIDAS, OASIS datasets, and the resultant network weights were compared across sites.

outperforms all ablated variants at each site ($p < 0.05$), except generally higher FID for $T_2 \rightarrow T_1$ in BRATS compared to variants, 'w/o PNA' that occasionally yields similar PSNR/SSIM and lower FID, and 'w/o STCA^b' that occasionally yields similar PSNR/FID. These results demonstrate the importance of STIN and STCA subblocks in PBs, partial network aggregation, and the mapper module. Second, we compared pFLSynth against variants where the site index was removed (w/o v^k), and the source-target configuration index was removed (w/o u_c) from the mapper. Performance in the variable task configuration is summarized in Table 7. pFLSynth consistently outperforms ablated variants across sites and tasks ($p < 0.05$), except for 'w/o u_c' that yields lower FID for $T_2 \rightarrow FR$ in OASIS. These results demonstrate that both site and source-target configuration information play a significant role in improving model specialization.

Next, we examined the utility of the PNA strategy in pFLSynth that only aggregates later generator stages. In theory, aggregation should be performed on network stages with more similar weights across sites, while stages with dissimilar weights should be kept local to help minimize loss of site-specific information (Ma et al., 2022). First, we assessed the similarity of weights for various generator stages by analyzing single-site pFLSynth$_{sing}$ models trained separately on each local dataset in the common task configuration. The mapper and PB blocks were removed from the architecture to avoid biases due to feature map modulations. At each stage, similarity between single-site models was taken as Spearman's correlation coefficient between the corresponding weight vectors. Fig. 8 displays the average similarity across sites as a function of generator stage. We find that later stages show substantially higher similarity across sites than earlier stages. The cumulative correlation is 0.84 in the later half of the convolutional backbone ($r6$-$d3$), as opposed to 0.16 in the first half ($e1$-$r5$). These findings indicate that PNA over later generator stages is a well-motivated strategy for improving site-specialization in federated synthesis models. We then conducted an ablation study to assess the selection of L_{cut} as $r5$ in pFLSynth based on validation performance. For this purpose, separate pFLSynth variants were trained while the generator cut-point was set as $L_{cut} = (r3, r4, r5, r6, r7)$. Performance metrics listed in Table 8 indicate that $L_{cut} = r5$ yields higher performance on average across sites, indicating that the proposed cut-point selection is favorable.

A practical concern in FL-based multi-institutional collaborations is reliability against delayed participation of a specific site or task (You et al., 2022). To examine this issue, spare digits were reserved in the site and source-target indices to code late joiners included halfway

Table 9
Effects of delayed participation. Results shown for the original pFLSynth with all sites/tasks included, pFLSynth with delayed site/task participation, and variant models ablated of site or task index.

			MIDAS			IXI
			T_1→T_2			T_2→PD
	pFLSynth (original)	PSNR ⇑	28.4 ± 0.6	pFLSynth (original)	PSNR ⇑	31.5 ± 1.0
		SSIM ⇑	92.0 ± 0.8		SSIM ⇑	97.4 ± 0.5
		FID ⇓	10.0		FID ⇓	21.6
Delayed participation	pFLSynth (delayed site)	PSNR ⇑	28.0 ± 0.5	pFLSynth (delayed task)	PSNR ⇑	31.0 ± 0.8
		SSIM ⇑	91.9 ± 1.2		SSIM ⇑	97.2 ± 0.7
		FID ⇓	10.6		FID ⇓	21.9
	w/o v_k (delayed site)	PSNR ⇑	23.0 ± 1.1	w/o u_c (delayed task)	PSNR ⇑	30.5 ± 0.9
		SSIM ⇑	68.8 ± 3.3		SSIM ⇑	96.6 ± 0.5
		FID ⇓	44.6		FID ⇓	24.7

Table 10
Performance for a held-out site. For site transfer, pFLSynth was trained on three sites (IXI, BRATS, OASIS) and tested on (MIDAS). The zero-shot variant was frozen after training. The fine-tuned variant was adapted on a compact set of MIDAS data. Results are also shown for the original pFLSynth model trained with all four sites included, and pFLSynth$_{sing}$ exclusively trained on the held-out MIDAS dataset.

			MIDAS	
			T_1→T_2	T_2→T_1
	pFLSynth (original)	PSNR ⇑	28.4 ± 0.6	26.2 ± 1.3
		SSIM ⇑	92.0 ± 0.8	86.5 ± 2.4
		FID ⇓	10.0	11.8
	pFLSynth$_{sing}$	PSNR ⇑	24.7 ± 0.5	22.2 ± 1.2
		SSIM ⇑	81.8 ± 0.9	70.4 ± 3.0
		FID ⇓	30.0	35.1
Site transfer	pFLSynth (zero-shot)	PSNR ⇑	18.7 ± 0.5	17.9 ± 0.8
		SSIM ⇑	64.9 ± 1.2	61.8 ± 1.8
		FID ⇓	168.1	155.4
	pFLSynth (fine-tuned)	PSNR ⇑	27.3 ± 0.5	24.5 ± 0.9
		SSIM ⇑	85.5 ± 1.1	79.0 ± 2.3
		FID ⇓	13.3	18.8

Table 11
Performance for a held-out task. For task transfer, pFLSynth was trained on all tasks except for T_1→T_2 in BRATS. The zero-shot variant was frozen after training. The fine-tuned variant was adapted on a compact set of BRATS data for T_1→T_2. Results are also shown for the original pFLSynth model trained with all tasks included, and pFLSynth$_{sing}$ trained for FR→T_2 and fine tuned for T_1→T_2 on BRATS.

			BRATS
			T_1→T_2
	pFLSynth (original)	PSNR ⇑	26.0 ± 0.9
		SSIM ⇑	93.5 ± 1.1
		FID ⇓	22.9
Task transfer	pFLSynth$_{sing}$ (fine-tuned)	PSNR ⇑	23.6 ± 1.1
		SSIM ⇑	87.4 ± 2.4
		FID ⇓	42.2
	pFLSynth (zero-shot)	PSNR ⇑	23.3 ± 1.1
		SSIM ⇑	88.4 ± 2.4
		FID ⇓	42.2
	pFLSynth (fine-tuned)	PSNR ⇑	24.8 ± 0.6
		SSIM ⇑	91.1 ± 1.5
		FID ⇓	22.3

during the training. For delayed site analysis, a single site was delayed in the common task configuration and pFLSynth was compared to a variant with ablated site index. For delayed task analysis, a single task was delayed in the variable task configuration and pFLSynth was compared to a variant with ablated source-target index. Performance metrics for the delayed site and task are listed in Table 9 (there were unsubstantial differences for non-delayed 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.

Another important concern pertains to the generalization of federated models to a new site or task. To examine generalization to a new site, pFLSynth was trained assuming a three-site FL setup (IXI, BRATS, OASIS) in the common task configuration and tested on the held-out site (MIDAS). A zero-shot variant was formed as the average of personalized pFLSynth models across training sites. A fine-tuned variant was formed by further training of the zero-shot variant on a compact subset of local MIDAS data (i.e., 1/4 th of the original MIDAS training set). For comparison, a single-site pFLSynth$_{sing}$ model was directly trained on the same subset of local data. Performance metrics for the held-out site are listed in Table 10. While the zero-shot variant performs suboptimally, the fine-tuned variant performs more competitively with the original pFLSynth trained in the four-site setup. In addition, the fine-tuned variant outperforms pFLSynth$_{sing}$ trained exclusively on local data from the held-out site. To examine generalization to a new task, pFLSynth was trained in the variable task configuration with all tasks except for T_1→T_2 in BRATS (note that T_1→T_2 was available in other sites), and then tested on the held-out task. A zero-shot variant was formed by simply probing the personalized pFLSynth model for BRATS with the source-target configuration index of the T_1→T_2 task. A fine-tuned variant was formed by further training of the zero-shot variant on a compact subset of local T_1→T_2 data from BRATS (1/4 th of the original BRATS training set for T_1→T_2). For comparison, a single-site pFLSynth$_{sing}$ model was trained for FR→T_2 on BRATS and then fine-tuned on the same subset of local T_1→T_2 data. Performance metrics for the held-out task are listed in Table 11. We find that the fine-tuned variant performs competitively with the original pFLSynth trained with all tasks included, while outperforming both the zero-shot variant and pFLSynth$_{sing}$. Taken together, these results suggest that pFLSynth shows a modest level of zero-shot generalization to new sites and tasks, and that these generalization abilities can be notably improved by transfer learning procedures. Furthermore, elevated performance of fine-tuned pFLSynth models over single-site models highlight a potential benefit of expanded training sets reached through multi-site collaborations in boosting model performance.

Lastly, we assessed the influence of PBs and PNA on avoiding possible information leakage during transfer of model parameters. Recent studies posit layer-wise measures to assess leakage in network models (Shwartz-Ziv and Tishby, 2017). Accordingly, we measured the similarity of activation maps in local synthesizer layers to assess the potential for leakage (Mo et al., 2020). A random set of 50 training source images were selected from each site and projected separately through all local generators. Similarity for a given source image was taken as Spearman's correlation coefficient between the activation maps it elicits in separate sites. Fig. 9 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 in the model architecture. Utilizing unshared PBs as in pFLSynth further reduces similarity in activation maps. These results suggest that both PBs and PNA in pFLSynth might help enhance reliability against information leakage.

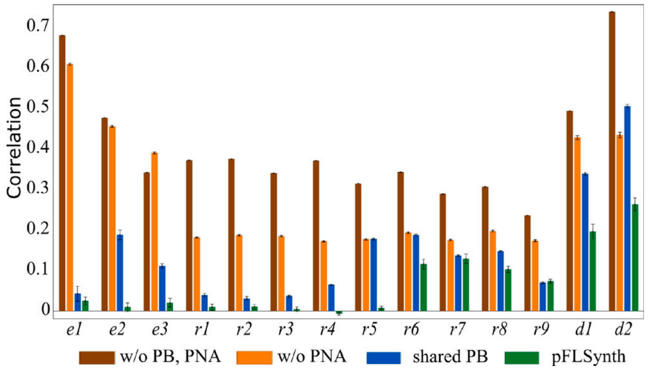


Fig. 9. Similarity of activation maps across generator stages, shown as mean±std of Spearman's correlation coefficient across sites. A set of training images were projected through the local generators at each site, and the resultant activation maps were compared across sites.

6. Discussion

Federated MRI synthesis has to operate under distributional heterogeneity in multi-site imaging data (Li et al., 2021). A recent study has considered FedAvg optimization of cycle-consistent models for MRI synthesis (Xie et al., 2022a). 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 multi-site synthesis model to each individual site and translation task. Experiments on multi-site MRI data demonstrate that pFLSynth offers on par performance to a centralized benchmark based on the proposed architecture, while outperforming other centralized and federated baselines. Therefore, our results suggest that pFLSynth can help improve generalizability and flexibility in multi-site collaborations by permitting training on imaging data from diverse sources. Here we reported experimental results from a single run of federated training, and from competing methods with moderate differences in model complexity. In unreported control analyses, we observed stable model performances with limited variability between independent training runs. In cases where training instabilities are suspected, robust training strategies for adversarial models could be adopted (Mescheder et al., 2018). We also observed that pFLSynth achieves similar performance improvements over baselines that were modified to have matching model complexity to pFLSynth. These observations suggest that our findings cannot be attributed to training variability in adversarial models or differences in model complexity.

Here we demonstrated the utility of synthetic MR images generated with pFLSynth by radiologically evaluating their visual quality and quantitatively evaluating their influence on a downstream segmentation task, in addition to direct assessments on synthesis performance. The latent spaces captured by pFLSynth through the personalized FL approach including PNA and PBs might also be useful in non-synthesis tasks such as image compression or classification. While a simple approach to form downstream models for such tasks is to use the latent representation at a specific stage of pFLSynth as model input, this can cause omission of information captured in the latent spaces at different stages. It remains important future work to systematically examine the utility of latent representations across various generator stages in pFLSynth in building downstream models.

The proposed pFLSynth method employs novel personalization blocks equipped with STIN and STCA subblocks to adjust statistics of feature maps across spatial and channel dimensions. Several medical imaging studies have recently proposed to incorporate normalization layers in synthesis models to lower model complexity (Song and Ye, 2021; Zhang et al., 2022; Denck et al., 2021). In Song and Ye (2021), Zhang et al. (2022), normalization was used to halve the number of parameters in a cycle-consistent model by sharing the generators in both synthesis directions. In Denck et al. (2021), normalization was used to synthesize images at different echo times (TE) for T_2-weighted MRI scans with a single model as opposed to separate models for each TE. Our approach differs significantly from these studies in that we utilize STIN subblocks for personalizing a federated synthesis model, as opposed to lowering model complexity. Our proposed STIN subblocks perform site- and task-specific normalization to cope with non-IID data distributions encountered in multi-site datasets. Furthermore, we also utilize STCA subblocks for site- and task-specific channel weighting to compensate for variations in the channel distribution of image features across multi-site datasets. These design elements enable the PBs to improve reliability of multi-site synthesis models against data heterogeneity.

An alternative approach to personalization for improving reliability would be to build multi-site synthesis models on enhanced training data. For instance, model reliability can be potentially boosted by performing training on a select set of less redundant, more complex and/or higher-quality source-target images within and across sites (Sharma and Hamarneh, 2020). Note that the select training set would still need to be of sufficient size to permit adequate training of deep-learning models, which assumes availability of a notably larger dataset at each site such that an effective selection can be performed. Consequently, this selection approach might suffer from suboptimal learning under data-limited settings that arise due to challenges in curating large-scale datasets. In contrast, the proposed personalized FL approach can leverage all available images without exclusion to increase the size of training sets, and instead attains reliability by modulating the network mapping as controlled by site and task indices. Thus, in data-limited settings, personalized FL might offer more efficient use of available training data (Sheller et al., 2020). It remains important future work to assess the relative benefits of select training sets versus personalization, and to examine their concurrent use in the context of federated MRI synthesis.

There are several limitations that might be addressed to further improve the performance and reliability of pFLSynth. A first line of improvement concerns the type of synthesis tasks that can be implemented. Here, we only considered supervised synthesis tasks where models were trained on paired datasets with registered source and target images from a matching set of subjects. Utilization of unpaired data can facilitate compilation of broader datasets for training substantially more complex models. In those cases, unsupervised (Wolterink et al., 2017; Ge et al., 2019; Xie et al., 2022b) or semi-supervised (Jin et al., 2019; Yurt et al., 2022) learning strategies can be adopted. Furthermore, here we only examined one-to-one synthesis tasks where a single source and a single target contrast were involved. For certain source-target configurations, such one-to-one contrast mapping will not be well posed, notably compromising task performance. For instance, MRI literature suggests that translations between endogenous contrasts (e.g., T_1, T_2) are often reasonably well posed, albeit translations involving exogenous contrasts (i.e., contrasts obtained by injection of external agents) might be ill posed (Lee et al., 2019). For this reason, T_{1c}-weighted images in BRATS with exogenous contrast were excluded from analysis in the current study. When information to synthesize the target modality is not sufficiently evident in a single source modality, multiple modalities that collectively carry the necessary information could be utilized and pFLSynth might be generalized to perform many-to-one mapping by increasing the number of input channels (Yurt et al., 2021; Sharma and Hamarneh, 2020).

A second line of improvement concerns the backbone models used to implement federated MRI synthesis. In the current study, pFLSynth was implemented using a GAN model built on a convolutional architecture. Despite their sensitivity to high-frequency details in medical images, GAN models can suffer from poor training stability and limited sample fidelity (Goodfellow et al., 2014). Implementing pFLSynth based on a diffusion-based model might help improve reliability and quality in multi-contrast MRI synthesis (Özbey et al., 2023). Furthermore, recent studies have reported benefits for training centralized models that leverage attention mechanisms such as transformers (Schlemper et al., 2019; Dalmaz et al., 2022; Liu et al., 2023). Given concerns on computational complexity and communication costs, it may be possible to boost the sensitivity of pFLSynth to long range context in MRI images by adopting efficient transformer architectures in the generator (Korkmaz et al., 2022; Jang and Hwang, 2022).

Another line of improvement concerns the FL procedures used for decentralized training of pFLSynth. Here, we primarily considered an inclusive FL setup where all sites and tasks contributed to the entire training process. When late-joining sites or tasks were present, we observed that successful learning can be achieved given a reasonable number of communication rounds after inclusion. An alternative is to use local mappers at each site to learn site- and task-specific latent variables without needing site/task indices. Certain scenarios might also involve inference at a new site or a new task that was held-out from training. Here we observed that the proposed synthesis model shows modest zero-shot generalization to a held-out site, albeit fine-tuning the synthesis model on a compact set of local data from the held-out site significantly boosts generalization. We also observed reasonable generalization performance when a particular task present in other sites during model training was transferred to a target site that excluded the task. While we did not examine generalization to a new task that was entirely absent from all training sites, our results imply that fine-tuning the synthesis model on a sufficient set of local data could help enhance synthesis performance. In the absence of fine-tuning data, test-time adaptation methods might be employed instead to improve generalization at the expense of added computational burden for inference (He et al., 2021). Here, model hyperparameters were selected based on average validation performance across sites. Such selection procedures might be impractical in scenarios including a large number of sites in the FL setup, or when a large number of hyperparameters are to be selected. In those cases, model performance might be improved by online tuning of site-specific hyperparameters via reinforcement-learning agents (Guo et al., 2022).

Here we primarily evaluated FL setups with comparable amounts of compute resources and training data at participating sites. As such, we employed synchronous FL where the server waits to receive locally-trained model parameters from all sites prior to aggregation (Xu et al., 2023). This approach offers reliability against heterogeneity in local training times due to cross-site differences in computing hardware and training set size. Yet, when cross-site differences in computation load are substantial, waiting for local updates from sites with lower compute resources or more data might cause notable delays in training of the global model. In such cases, asynchronous FL could be adopted by prescribing a fixed duration for each communication round, and aggregating local models from the subset of sites that have finished their computations in each round (Xu et al., 2023). Asynchronous FL can enhance training efficiency at the expense of potential biases in the global model towards sites that can more rapidly compute local model updates. Thus, advanced aggregation algorithms that adaptively weight contributions from individual sites might be necessary to alleviate biases during asynchronous FL (Nguyen et al., 2022).

FL avoids transfer of imaging data to mitigate patient privacy risks. Yet, inference attacks might leak information about training data from model parameters (Kaissis et al., 2020). Here, we considered an FL setup where the 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 (Han et al., 2020). Moreover, pFLSynth leverages partial network aggregation such that only convolutional blocks in later stages of the generator are communicated. Note that it is highly challenging to conduct an inference attack based on a partial network that is nearly halved in size. Nevertheless, potential risks can be further minimized by adopting differentially private training (Xie et al., 2022a; Ziller et al., 2021), or by extending the size and diversity of the training datasets to implicitly improve privacy (Feng et al., 2021a). Future studies are warranted to systematically examine the privacy properties of FL-based methods in multi-contrast MRI synthesis.

The primary application for pFLSynth is sequence imputation in multi-contrast MRI protocols. Running prolonged exams with many MRI sequences is challenging due to economic/labor costs and motion artifacts in patients with difficulty to remain still (Dar et al., 2019). MRI synthesis models can enhance downstream analyses by recovering missing sequences from a subset of successfully acquired sequences (Iglesias et al., 2013). Such imputed protocols can help improve time- and cost-efficiency of MRI exams, or facilitate enrollment of larger patient cohorts in clinical studies. In this context, pFLSynth can enable collaborative training of multi-site models that reliably generalize across sites while maintaining site-specific features. Our quantitative evaluations indicate that pFLSynth offers significant performance benefits for both common and variable task configurations across sites. Results in the latter scenario indicate that, even when separate sites prescribe different synthesis tasks, the aggregated components of a federated synthesis model that are shared across sites can significantly benefit from learning on a larger and more diverse training set with partially overlapping contrasts among sites. The preliminary radiological evaluations reported here indicate that pFLSynth attains high visual similarity to ground-truth target contrasts, highlighting its potential to produce diagnostically-valuable images. Yet, future work is warranted to validate the utility of pFLSynth on large patient cohorts and in subsequent image analysis tasks (Huo et al., 2018; Yang et al., 2019).

Another potential application of pFLSynth is domain adaptation of downstream segmentation or classification models across MRI contrasts and imaging sites (Wu and Zhuang, 2021). When the amount of labeled data to train a downstream model is limited in a primary domain, the model can first be trained in a secondary domain with ample data and then transferred to the primary domain. By translating test images from the primary to the secondary domain, pFLSynth might improve the performance of the transferred model. Finally, pFLSynth might also be employed for translation tasks involving other modalities such as CT or PET (Huynh et al., 2016; Zhao et al., 2017; Hu et al., 2020; Wei et al., 2020), and for inverse problems in imaging such as reconstruction or super-resolution (Gungor et al., 2023; Güngör et al., 2022).

7. 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. Comprehensive experiments on multi-site datasets with common and variable task configurations were presented to demonstrate the benefits of pFLSynth over prior federated methods for brain image synthesis. Improved performance under implicit and explicit data heterogeneity renders pFLSynth a promising candidate for multi-institutional collaborations in multi-contrast MRI synthesis.

CRediT authorship contribution statement

Onat Dalmaz: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Muhammad U. Mirza:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gokberk Elmas:** Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Muzaffer Ozbey:** Writing – original draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Salman U.H. Dar:** Writing – original draft, Visualization, Validation, Software,

Investigation, Formal analysis, Data curation. **Emir Ceyani:** Writing – review & editing, Visualization, Validation, Investigation, Formal analysis. **Kader K. Oguz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis. **Salman Avestimehr:** Writing – original draft, Visualization, Validation, Investigation, Formal analysis. **Tolga Çukur:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103121.

References

Adam, A., Dixon, A., Gillard, J., Schaefer-Prokop, C., Grainger, R., Allison, D., 2014. Grainger & Allison's Diagnostic Radiology. Elsevier.

Altman, D.G., Bland, J.M., 2007. Missing data. BMJ 334 (7590), 424.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B., 2020. MedGAN: Medical image translation using GANs. Comput. Med. Imaging Graph. 79, 101684.

Atlas, S., 2009. Magnetic Resonance Imaging of the Brain and Spine. In: LWW medical book collection, (1. c.), Wolters Kluwer Health/Lippincott Williams & Wilkins.

Bakas, S., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data 4.

Beers, A., Brown, J., Chang, K., Campbell, J., Ostmo, S., Chiang, M., K.-Cramer, J., 2018. High-resolution medical image synthesis using progressively grown generative adversarial networks. arXiv:1805.03144.

Billot, B., Colin, Y., Das, S., Iglesias, J.E., 2023a. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. Proc. Natl. Acad. Sci. 120 (9), 1–10.

Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., 2023b. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. Med. Image Anal. 86, 102789.

Bowles, C., Qin, C., Ledig, C., Guerrero, R., Gunn, R., Hammers, A., Sakka, E., Dickie, D., Hernández, M., Royle, N., Wardlaw, J., Rhodius-Meester, H., Tijms, B., Lemstra, A., Flier, W., Barkhof, F., Scheltens, P., Rueckert, D., 2016. Pseudo-healthy image synthesis for white matter lesion segmentation. In: Simul. Synth. Med. Imaging. pp. 87–96.

Buda, M., Saha, A., Mazurowski, M.A., 2019. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput. Biol. Med. 109, 218–225.

Bullitt, E., Zeng, D., Gerig, G., Aylward, S., Joshi, S., Smith, J., Lin, W., Ewend, M., 2005. Vessel tortuosity and brain tumor malignancy. Acad. Radiol. 12, 1232–1240.

Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A., 2017. Adversarial image synthesis for unpaired multi-modal cardiac data. In: Simul. Synth. Med. Imaging. pp. 3–13.

Che, S., Kong, Z., Peng, H., Sun, L., Leow, A., Chen, Y., He, L., 2022. Federated multi-view learning for private medical data integration and analysis. ACM Trans. Intell. Syst. Technol. 13 (4).

Cheng, G., Chadha, K.N., Duchi, J.C., 2021. Fine-tuning is fine in federated learning. arXiv:2018.07313.

Clark, L.T., Watkins, L., Pina, I.L., Elmer, M., Akinboboye, O., Gorham, M., Jamerson, B., McCullough, C., Pierre, C., Polis, A.B., Puckrein, G., Regnante, J.M., 2019. Increasing diversity in clinical trials: Overcoming critical barriers. Cur. Prob. Cardiol. 44 (5), 148–172.

Cordier, N., Delingette, H., Le, M., Ayache, N., 2016. Extended modality propagation: Image synthesis of pathological cases. IEEE Trans. Med. Imaging 35 (12), 2598–2608.

Dalmaz, O., Yurt, M., Çukur, T., 2022. ResViT: Residual vision transformers for multi-modal medical image synthesis. IEEE Trans. Med. Imaging 41 (10), 2598–2614.

Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Çukur, T., 2019. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE Trans. Med. Imaging 38 (10), 2375–2388.

Denck, J., Guehring, J., Maier, A., Rothgang, E., 2021. MR-contrast-aware image-to-image translations with generative adversarial networks. Int. J. Comput. Assist. Rad. Surge. 16.

Ellison, D., Love, S., Chimelli, L., Harding, B., Lowe, J., Vinters, H., Brandner, S., Yong, W., 2012. Neuropathology: A Reference Text of CNS Pathology. Elsevier.

Elmas, G., Dar, S.U., Korkmaz, Y., Ceyani, E., Susam, B., Özbey, M., Avestimehr, S., Çukur, T., 2022. Federated learning of generative image priors for MRI reconstruction. IEEE Trans. Med. Imaging 42 (7), 1996–2009.

Fallah, A., Mokhtari, A., Ozdaglar, A., 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In: NeurIPS. pp. 3557–3568.

Feng, Q., Guo, C., Benitez-Quiroz, F., Martinez, A., 2021a. When do GANs replicate? On the choice of dataset size. In: ICCV. pp. 6701–6710.

Feng, C.-M., Yan, Y., Fu, H., Xu, Y., Shao, L., 2021b. Specificity-preserving federated learning for MR image reconstruction. arXiv:2112.05752.

Ge, Y., Wei, D., Xue, Z., Wang, Q., Zhou, X., Zhan, Y., Liao, S., 2019. Unpaired MR to CT synthesis with explicit structural constrained adversarial learning. In: IEEE ISBI. pp. 1096–1099.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. In: Adv. Neural Inf. Process. Syst., vol. 24.

Güngör, A., Askin, B., Soydan, D.A., Saritas, E.U., ş Top, C.B., Çukur, T., 2022. TranSMS: Transformers for super-resolution calibration in magnetic particle imaging. IEEE Trans. Med. Imaging 41 (12), 3562–3574.

Gungor, A., Dar, S.U., Ozturk, S., Korkmaz, Y., Elmas, G., Ozbey, M., Çukur, T., 2023. Adaptive diffusion priors for accelerated MRI reconstruction. Med. Image Anal. 88, 102872.

Guo, P., Wang, P., Yasarla, R., Zhou, J., Patel, V.M., Jiang, S., 2021a. Anatomic and molecular MR image synthesis using confidence guided CNNs. IEEE Trans. Med. Imaging 40 (10), 2832–2844.

Guo, P., Wang, P., Zhou, J., Jiang, S., Patel, V.M., 2021b. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. arXiv:2103.02148.

Guo, P., Yang, D., Hatamizadeh, A., Xu, A., Xu, Z., Li, W., Zhao, C., Xu, D., Harmon, S., Turkbey, E., Turkbey, B., Wood, B., Patella, F., Stellato, E., Carrafiello, G., Patel, V.M., Roth, H.R., 2022. Auto-FedRL: Federated hyperparameter optimization for multi-institutional medical image segmentation. In: ECCV. pp. 437–455.

Halligan, S., Kenis, S.F., Abeyakoon, O., Plumb, A.A.O., Mallett, S., 2020. How to avoid describing your radiological research study incorrectly. Eur. Radiol. 30 (8), 4648–4655.

Han, T., Nebelung, S., Haarburger, C., Horst, N., Reinartz, S., Merhof, D., Kiessling, F., Schulz, V., Truhn, D., 2020. Breaking medical data sharing boundaries by using synthesized radiographs. Sci. Adv. 6 (49), eabb7973.

He, Y., Carass, A., Zuo, L., Dewey, B.E., Prince, J.L., 2021. Autoencoder based self-supervised test-time adaptation for medical image analysis. Med. Image Anal. 72, 102136.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR. pp. 770–778.

Hu, S., Shen, Y., Wang, S., Lei, B., 2020. Brain MR to PET synthesis via bidirectional generative adversarial network. In: Med Imag Comput Comput Assist Intern. pp. 698–707.

Huang, X., Belongie, S.J., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. arXiv:1703.06868.

Huang, Y., Shao, L., Frangi, A.F., 2018. Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. IEEE Trans. Med. Imaging 37 (3), 815–827.

Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R.G., Landman, B.A., 2018. Adversarial synthesis learning enables segmentation without target modality ground truth. In: IEEE ISBI. pp. 1217–1220.

Huynh, T., Gao, Y., Kang, J., Wang, L., Zhang, P., Lian, J., Shen, D., 2016. Estimating ct image from mri data using structured random forest and auto-context model. IEEE Trans. Med. Imaging 35 (1), 174–183.

Iglesias, J.E., Konukoglu, E., Zikic, D., Glocker, B., Van Leemput, K., Fischl, B., 2013. Is synthesizing mri contrast useful for inter-modality analysis? In: MICCAI. pp. 631–638.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. CVPR 1125–1134.

Jang, J., Hwang, D., 2022. M3T: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In: IEEE CVPR. pp. 20686–20697.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.

Jin, C.-B., Kim, H., Liu, M., Jung, W., Joo, S., Park, E., Ahn, Y.S., Han, I.H., Lee, J.I., Cui, X., 2019. Deep CT to MR synthesis using paired and unpaired data. Sensors 19 (10), 2361.

Jog, A., Carass, A., Roy, S., Pham, D.L., Prince, J.L., 2017. Random forest regression for magnetic resonance image synthesis. Med. Image Anal. 35, 475–488.

Joyce, T., Chartsias, A., Tsaftaris, S.A., 2017. Robust multi-modal mr image synthesis. In: MICCAI. pp. 347–355.

Kaissis, G.A., Makowski, M.R., Rüeckert, D., Braren, R.F., 2020. Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. 2 (6), 305–311.

Knoll, F., Zbontar, J., Sriram, A., Muckley, M.J., Bruno, M., Defazio, A., Parente, M., Geras, K.J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzalv, M., Romero, A., Rabbat, M., Vincent, P., Pinkerton, J., Wang, D., Yakubova, N., Owens, E., Zitnick, C.L., Recht, M.P., Sodickson, D.K., Lui, Y.W., 2020. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. Rad. Artif. Intell. 2 (1), e190007.

Korkmaz, Y., Dar, S.U., Yurt, M., Özbey, M., Çukur, T., 2022. Unsupervised MRI reconstruction via zero-shot learned adversarial transformers. IEEE Trans. Med. Imaging 41 (7), 1747–1763.

Krupa, K., Bekiesińska-Figatowska, M., 2015. Artifacts in magnetic resonance imaging. Pol. J. Radiol. 80, 93–106.

LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., Raichle, M.E., Cruchaga, C., Marcus, D., 2019. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. medRxiv http://dx.doi.org/10.1101/2019.12.13.19014902.

Lan, H., the Alzheimer Disease Neuroimaging Initiative, Toga, A.W., Sepehrband, F., 2021. Three-dimensional self-attention conditional gan with spectral normalization for multimodal neuroimaging synthesis. Magn. Reson. Med. 86 (3), 1718–1733.

Lee, D., Kim, J., Moon, W.-J., Ye, J.C., 2019. CollaGAN: Collaborative GAN for missing image data imputation. In: CVPR. pp. 2487–2496.

Lei, Y., Fu, Y., Mao, H., Curran, W.J., Liu, T., Yang, X., 2020. Multi-modality MRI arbitrary transformation using unified generative adversarial networks. In: SPIE Med. Imaging, vol. 11313, 1131303.

Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S., 2020a. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med. Image Anal. 65, 101765.

Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q., 2021. FedBN: Federated learning on non-IID features via local batch normalization. In: ICLR.

Li, X., Li, M., Yan, P., Li, G., Jiang, Y., Luo, H., Yin, S., 2023. Deep learning attention mechanism in medical image analysis: Basics and beyonds. Int. J. Net. Dyn. Intell. 2 (1), 93–116.

Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., Feng, A., 2019a. Privacy-preserving federated brain tumour segmentation. In: Mach. Learn. Med. Imaging. pp. 133–141.

Li, H., Paetzold, J.C., Sekuboyina, A., Kofler, F., Zhang, J., Kirschke, J.S., Wiestler, B., Menze, B., 2019b. DiamondGAN: Unified multi-modal generative adversarial networks for MRI sequences synthesis. In: MICCAI. pp. 795–803.

Li, T., Sahu, A.K., Talwalkar, A.S., Smith, V., 2020b. Federated learning: Challenges, methods, and future directions. IEEE Signal. Process. Mag. 37, 50–60.

Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P., 2021. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: CVPR. pp. 1013–1023.

Liu, J., Pasumarthi, S., Duffy, B., Gong, E., Datta, K., Zaharchuk, G., 2023. One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. IEEE Trans. Med. Imaging 1. http://dx.doi.org/10.1109/TMI.2023.3261707.

Luo, Y., Nie, D., Zhan, B., Li, Z., Wu, X., Zhou, J., Wang, Y., Shen, D., 2021. Edge-preserving MRI image synthesis via adversarial network with iterative multi-scale fusion. Neurocomputing 452, 63–77.

Ma, X., Zhang, J., Guo, S., Xu, W., 2022. Layer-wised model aggregation for personalized federated learning. In: IEEE CVPR. pp. 10082–10091.

Mansour, Y., Mohri, M., Ro, J., Suresh, A.T., 2020. Three approaches for personalization with applications to federated learning. arXiv:2002.10619.

McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data. In: AISTATS.

Menze, B.H., et al., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024.

Mescheder, L., Geiger, A., Nowozin, S., 2018. Which training methods for GANs do actually converge? In: ICML. pp. 3481–3490.

Mo, F., Borovykh, A., Malekzadeh, M., Haddadi, H., Demetriou, S., 2020. Layer-wise characterization of latent information leakage in federated learning. arXiv:2010.08762.

Nguyen, J., Malik, K., Zhan, H., Yousefpour, A., Rabbat, M., Malek, M., Huba, D., 2022. Federated learning with buffered asynchronous aggregation. In: Int. Conf. Artif. Intell. Stat., vol. 151, PMLR, pp. 3581–3607.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical image synthesis with deep convolutional adversarial networks. IEEE Trans. Biomed. Eng. 65 (12), 2720–2730.

Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Ozturk, S., Gungor, A., Çukur, T., 2023. Unsupervised medical image translation with adversarial diffusion models. IEEE Trans. Med. Imaging 42 (12), 3524–3539.

Park, S., Kim, G., Kim, J., Kim, B., Ye, J.C., 2021. Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis. In: NeurIPS.

Rasouli, M., Sun, T., Rajagopal, R., 2020. FedGAN: Federated generative adversarial networks for distributed data. arXiv:2006.07228.

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R.M., Trask, A., Xu, D., Baust, M., Cardoso, M.J., 2020. The future of digital health with federated learning. npj Dig. Med. 3 (1), 119.

Rodney, R.A., 2021. The attainment of patient diversity in clinical trials: Race, ethnicity, genetics. Am. J. Med. 134 (12), 1440–1441.

Roth, H.R., Yang, D., Li, W., Myronenko, A., Zhu, W., Xu, Z., Wang, X., Xu, D., 2021. Federated whole prostate segmentation in MRI with personalized neural architectures. In: MICCAI. pp. 357–366.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to Leverage Salient Regions in medical images. Med. Image Anal. 53, 197–207.

Sevetlidis, V., Giuffrida, M.V., Tsaftaris, S.A., 2016. Whole image synthesis using a deep encoder-decoder network. In: Simul. Synth. Med. Imaging. pp. 127–137.

Sharma, A., Hamarneh, G., 2020. Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. IEEE Trans. Med. Imaging 39 (4), 1170–1183.

Sheller, M., Edwards, B., Reina, G., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R., Bakas, S., 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. 10.

Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S., 2019. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In: MICCAI. pp. 92–104.

Shwartz-Ziv, R., Tishby, N., 2017. Opening the black box of deep neural networks via information. arXiv:1703.00810.

Singh, A., Horng, H., Chitalia, R., Roshkovan, L., Katz, S.I., Noël, P., Shinohara, R.T., Kontos, D., 2022. Resampling and harmonization for mitigation of heterogeneity in image parameters of baseline scans. Sci. Rep. 12 (1), 21505.

Song, J., Ye, J.C., 2021. Federated CycleGAN for privacy-preserving image-to-image translation. arXiv:2106.09246.

Tan, A.Z., Yu, H., Cui, L., Yang, Q., 2023. Towards personalized federated learning. IEEE Trans. Neural Netw. Learn. Syst. 34 (12), 9587–9603.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2017. Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022.

Van Nguyen, H., Zhou, K., Vemulapalli, R., 2015. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: MICCAI. pp. 677–684.

Vemulapalli, R., Van Nguyen, H., Zhou, S.K., 2015. Unsupervised cross-modal synthesis of subject-specific scans. In: ICCV. pp. 630–638.

Wang, G., Gong, E., Banerjee, S., Martin, D., Tong, E., Choi, J., Chen, H., Wintermark, M., Pauly, J.M., Zaharchuk, G., 2020a. Synthesize high-quality multi-contrast magnetic resonance imaging from multi-echo acquisition using multi-task deep generative model. IEEE Trans. Med. Imaging 39 (10), 3089–3099.

Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J., 2020b. Deep multimodal fusion by channel exchanging. In: NeurIPS.

Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Rad. Artif. Intelli. 5 (5), e230024.

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Colliot, O., Stankoff, B., Ayache, N., 2019. Fluid-attenuated inversion recovery mri synthesis from multisequence mri using three-dimensional fully convolutional networks for multiple sclerosis. J. Med. Imaging 6 (1), 014005.

Wei, W., Poirion, E., Bodini, B., Tonietto, M., Durrleman, S., Colliot, O., Stankoff, B., Ayache, N., 2020. Predicting PET-derived myelin content from multisequence MRI for individual longitudinal analysis in multiple sclerosis. NeuroImage 223, 117308.

Wolterink, J., Dinkla, A.M., Savenije, M., Seevinck, P., Berg, C., Isgum, I., 2017. Deep MR to CT synthesis using unpaired data. In: Simul. Synth. Med. Imaging. pp. 14–23.

Wu, Y., He, K., 2018. Group normalization. arXiv:1803.08494.

Wu, F., Zhuang, X., 2021. Unsupervised domain adaptation with variational approximation for cardiac segmentation. IEEE Trans. Med. Imaging 40 (12), 3555–3567.

Xie, G., Wang, J., Huang, Y., Li, Y., Zheng, Y., Zheng, F., Jin, Y., 2022a. FedMed-GAN: Federated domain translation on unsupervised cross-modality brain image synthesis. arXiv:2201.08953.

Xie, G., Wang, J., Huang, Y., Zheng, Y., Zheng, F., Jin, Y., 2022b. FedMed-ATL: Misaligned unpaired brain image synthesis via affine transform loss. arXiv:2201.12589.

Xu, C., Qu, Y., Xiang, Y., Gao, L., 2023. Asynchronous federated learning on heterogeneous devices: A survey. Comput. Sci. Rev. 50, 100595.

Yan, Z., Wicaksana, J., Wang, Z., Yang, X., Cheng, K.T., 2021. Variation-aware federated learning with multi-source decentralized medical image data. IEEE J. Biomed. Health Inf. 25 (7), 2615–2628.

Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S., 2019. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In: MICCAI. pp. 255–263.

Yang, H., Lu, X., Wang, S.-H., Lu, Z., Yao, J., Jiang, Y., Qian, P., 2021. Synthesizing multi-contrast mr images via novel 3D conditional variational auto-encoding gan. Mob. Netw. Appl. 26, 1–10.

You, C., Xiang, J., Su, K., Zhang, X., Dong, S., Onofrey, J., Staib, L., Duncan, J.S., 2022. Incremental learning meets transfer learning: Application to multi-site prostate MRI segmentation. In: MICCAI DeCAF. pp. 3–16.

Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J., Bourgeat, P., 2019. Ea-GANs: Edge-aware generative adversarial networks for cross-modality MR image synthesis. IEEE Trans. Med. Imaging 38 (7), 1750–1762.

Yurt, M., Dar, S.U., Erdem, A., Erdem, E., Oguz, K.K., Çukur, T., 2021. mustGAN: multi-stream generative adversarial networks for MR image synthesis. Med. Image Anal. 70, 101944.

Yurt, M., Dar, S.U.H., Özbey, M., Tınaz, B., Oğuz, K.K., Çukur, T., 2022. Semi-supervised learning of mutually accelerated MRI synthesis without fully-sampled ground truths. IEEE Trans. Med. Imaging 41 (12), 3895–3906.

Zhan, B., Di Li, Wang, Y., Ma, Z., Wu, X., Zhou, J., Zhou, L., 2021. LR-cGAN: Latent representation based conditional generative adversarial network for multi-modality MRI synthesis. Biomed. Signal Process. Control 66, 102457.

Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2019. Self-attention generative adversarial networks. In: ICML. pp. 7354–7363.

Zhang, H., Li, H., Dillman, J., Parikh, N., He, L., 2022. Multi-contrast MRI image synthesis using switchable cycle-consistent generative adversarial networks. Diagnostics 12, 816.

Zhao, C., Carass, A., Lee, J., He, Y., Prince, J.L., 2017. Whole brain segmentation and labeling from ct using synthetic mr images. In: Mach. Learn. Med. Imaging. pp. 291–298.

Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L., 2020. Hi-net: Hybrid-fusion network for multi-modal mr image synthesis. IEEE Trans. Med. Imaging 39 (9), 2772–2781.

Ziller, A., Usynin, D., Knolle, M., Hammernik, K., Rueckert, D., Kaissis, G., 2021. Complex-valued deep learning with differential privacy. arXiv:2110.03478.