Multi-institutional Collaborations for Improving Deep Learning-based Magnetic Resonance Image Reconstruction Using Federated Learning
——*Supplementary Material*——

Pengfei Guo Puyang Wang Jinyuan Zhou Shanshan Jiang Vishal M. Patel
Johns Hopkins University
{pguo4,pwang47}@jhu.edu, {jzhou2,sjiang21}@jhmi.edu, vpatel36@jhu.edu

1. Supplementary Introduction

In this supplementary document, we first provide supplementary method including details of network architecture and addition schematics of FL-MRCM. Then, we present additional experimental results including experiments on different acceleration factors, more qualitative results, and ablation studies.

2. Supplementary Method

We use a U-Net [4] style encoder-decoder architecture for the reconstruction networks. Table 1 shows the details of each block in encoder and decoder in our reconstruction network. Note Conv and ConvTranspose denote the 2D convolution and 2D transposed convolution operator, respectively. The encoder networks can be described as follows:

ConvBlock(1,32)-AvgPool(2,2)-ConvBlock(32,64)-AvgPool(2,2)-ConvBlock(64,128)-AvgPool(2,2)-ConvBlock(128,256)-AvgPool(2,2)-ConvBlock(256,512),

where AvgPool(2,2) represents 2D average pooling with kernel size of 2 and stride size of 2 and other network modules are express by (in-channel, out-channel). Then, feature maps are projected to latent space as the input of domain identifiers by an adaptive average pooling layer with outshape $512 \times 2 \times 2$. The decoder networks can be expressed as follows:

Upsample(512,256)-ConvBlock(512,256)-Upsample(256,128)-ConvBlock(256,128)–Upsample(128,64)-ConvBlock(128,64)–Upsample(64,32)-ConvBlock(64,32)-Conv(32,1).

The domain identifier consists of two fully connected layers as follow:

FC(2048,256)-LeakyRelu(0.2)-FC(256,2),

where LeakyRelu(0.2) represents the LeakyRelu activation with negative slope of 0.2.

Figure 1 shows a global view of proposed FL-MRCM for multi-institutional collaborations in MR image reconstruction task. For the target site, the decoder part (green block) and the ground truth image are transparent, since they might be not involved during the training. As discussed in Section 4 of the main manuscript, there is not fully-sampled data for training in Scenario 1.

3. Additional Experimental Results

The ablation study about the effectiveness of proposed cross-site modeling is demonstrated by a set of the comparisons between FL-MR and FL-MRCM under the setting of federated learning. Furthermore, we also conduct a detailed ablation study to analyze the effectiveness of proposed cross-site modeling without federated learning framework for T_1-weighted images. In this case, we obtain a trained model from one of available sites and evaluate the its performance on another institution to observe the gain purely contributed by cross-site modeling in Table 2. We present the experiment results when the acceleration factor is set to 8 in Table 4. Similar with results of acceleration factor of 4 in main manuscript, our proposed FL-MR exhibits better generalization and clearly outperforms other privacy-preserving alternative strategies. FL-MRCM outperforms FL-MR in each dataset by addressing the domain shift issue.

Table 5 is a extended version of Table 2 in the main manuscript. We additionally compare the performance of

Table 1. Configuration of Blocks in FL-MRCM

Block	Layer	Kernel size	Stride	Padding
ConvBlock	Conv	3	1	1
	InstanceNorm	-	-	-
	LeakyReLu	-	-	-
	Conv	3	1	1
	InstanceNorm	-	-	-
	LeakyReLu	-	-	-
UpSample	ConvTranspose	2	2	0
	InstanceNorm	-	-	-
	LeakyReLu	-	-	-

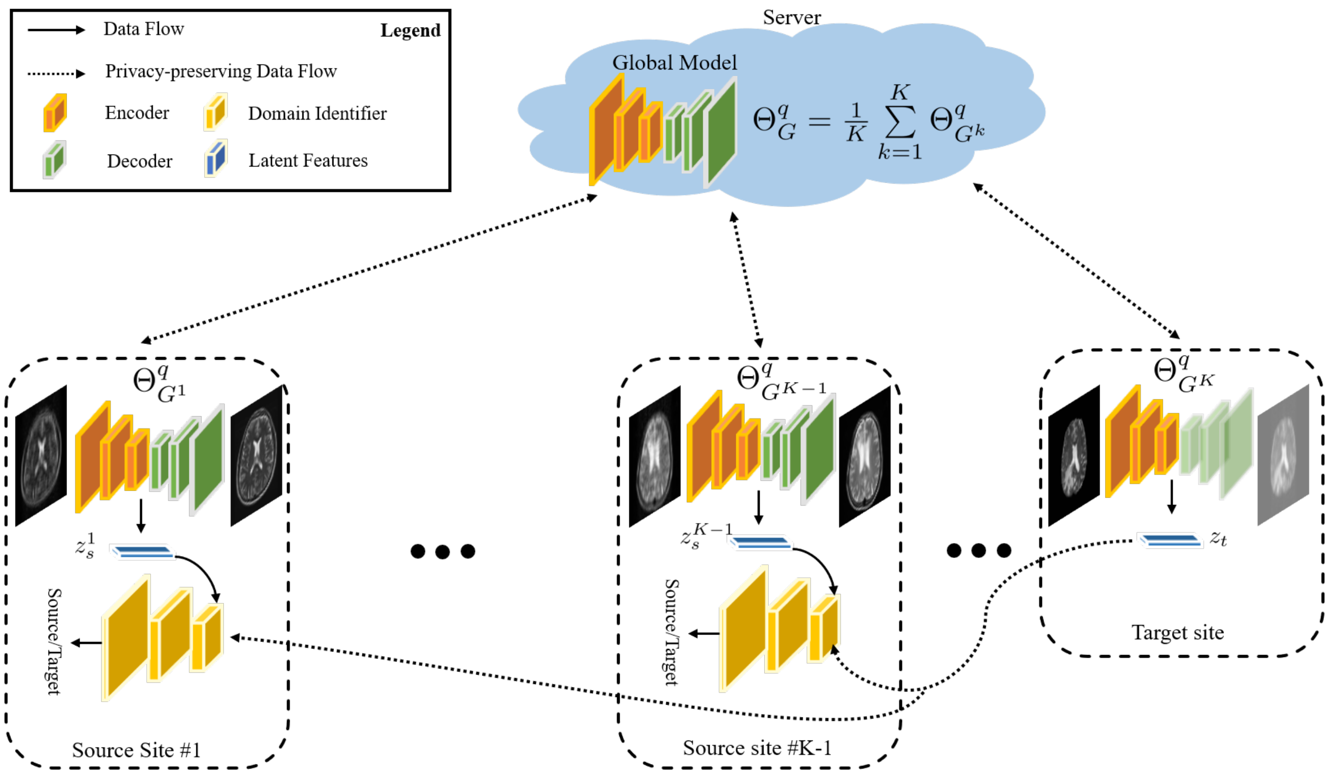

Figure 1. The overview of the proposed FL-MRCM framework. Through several rounds of communication between data centers and server, the collaboratively trained global model parameterized by Θ_G^q can be obtained in a data privacy-preserving manner.

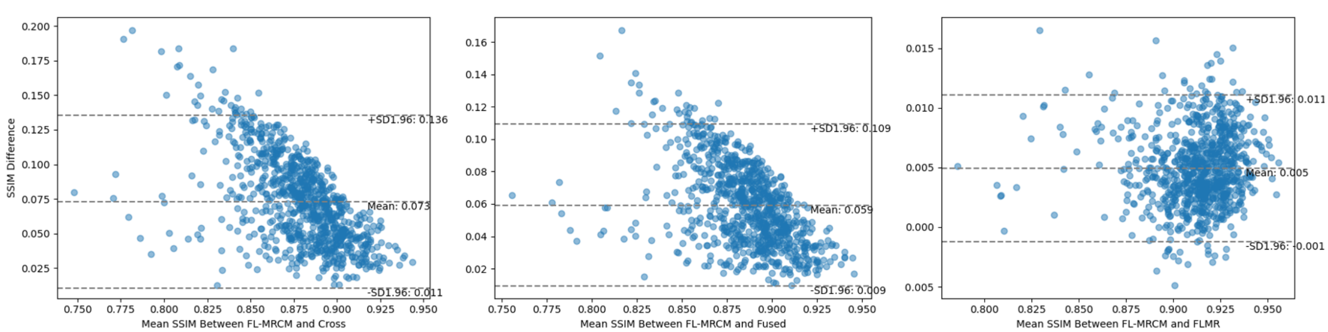

Figure 2. Bland–Altman plot corresponding to the fastMRI dataset between FL-MRCM and other methods in Scenario 1.

the proposed framework with models pre-trained with data from a single data center and then fine-tuned with data from target data center. In this case, we obtain a trained model from one of the institutions, then we transfer the pre-trained weights to the target site and fine-tune the pre-trained model by the training data of the target site, which will not compromises the data sharing regulations. We denote this set of experiments as **Transfer** in Table 5. The reported results suggest that pre-trained on a large dataset (e.g., the F dataset) can improve the performance but the multi-institutional collaboration is still a better option if multiple datasets are available.

Figure 3 shows the qualitative performance of different methods on T_1 and T_2-weighted images from four datasets in Scenario 2. It can be observed that the proposed FL-MRCM method yields reconstructed images with remarkable visual similarity to the reference images compared to the other alternatives (see the last column of each sub-figure in Fig. 3) in four datasets with diverse characteristics.

To investigate the performance improvement of the proposed FL-MRCM, we conduct t-test based on the SSIM of the reconstructed images between FL-MRCM and other methods. Averaged p values of each group of experiments in two scenarios are presented in Table 3. A p value less than 0.05 is usually considered as statistically significant. The reported performance of FL-MRCM satisfies this criterion. To further demonstrate the performance of the proposed FL-MRCM, we show an example of Bland–Altman

Table 2. Quantitative ablation study of proposed cross-site modeling on T_1-weighted images. For experiments with cross-site modeling, the target site is the institution that provides testing data.

Data Centers (Institutions)		w/o Cross-site Modeling				w/ Cross-site Modeling			
		SSIM	PSNR	Average		SSIM	PSNR	Average	
Train	Test			SSIM	PSNR			SSIM	PSNR
B	F	0.7694	28.61			0.7987	29.53		
B	H	0.5188	25.07	0.7222	27.93	0.5350	25.08	**0.7399**	**28.42**
B	I	0.8785	30.10			0.8859	30.65		
F	B	0.9016	34.65			0.9158	35.13		
F	H	0.8402	28.52	0.8840	31.44	0.8603	28.83	**0.8978**	**31.79**
F	I	0.9102	31.16			0.9172	31.42		
H	B	0.6670	29.12			0.7256	31.54		
H	F	0.8571	31.82	0.7736	30.03	0.8938	32.59	**0.8292**	**31.40**
H	I	0.7968	29.16			0.8681	30.07		
I	B	0.8795	33.76			0.9310	35.03		
I	F	0.8417	31.18	0.7831	30.68	0.8803	31.81	**0.8522**	**31.58**
I	H	0.6281	27.09			0.7454	27.89		

plot for fastMRI (the largest dataset) in Fig. 2. The y axis represents the SSIM difference of the reconstructed images between FL-MRCM and other methods. We can observe that most points lie in the positive range, which implies that FL-MRCM exhibits better reconstruction performance on most subjects.

Table 3. The p values of t-test among different methods in two scenarios.

Scenario 1			Scenario 2		
Method	T_1-weighted	T_2-weighted	Method	T_1-weighted	T_2-weighted
Cross	4.11×10^{-27}	2.76×10^{-07}	Single	3.20×10^{-02}	7.92×10^{-03}
Fused	9.51×10^{-20}	1.12×10^{-15}	FL-MR	2.20×10^{-02}	4.34×10^{-07}
FL-MR	5.28×10^{-05}	6.47×10^{-03}	FL-MRCM	-	-
FL-MRCM	-	-	-	-	-

While our proposed method yields better performance, there are several limitations in our current study. First, experiments are conducted on the same sequences (e.g., T_1 and T_2) with the Cartesian undersampling. Although T_1 and T_2 are widely used sequences in clinical practice and Cartesian undersampling is usually adapted by compressed sensing, this might limit the applicability of our approach. The proposed method is inherently compatible with different kinds of sequences and undersampling. We will explore this direction in our future work. Second, experiments are based on simulated acquisition (starting from fully-sampled k-space and simulating acceleration). Further verification of accelerated acquisition on actual scanners will make this study more persuasive.

References

[1] brain development.org. 4

[2] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020. 4

[3] Bjoern H Menze et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 4

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

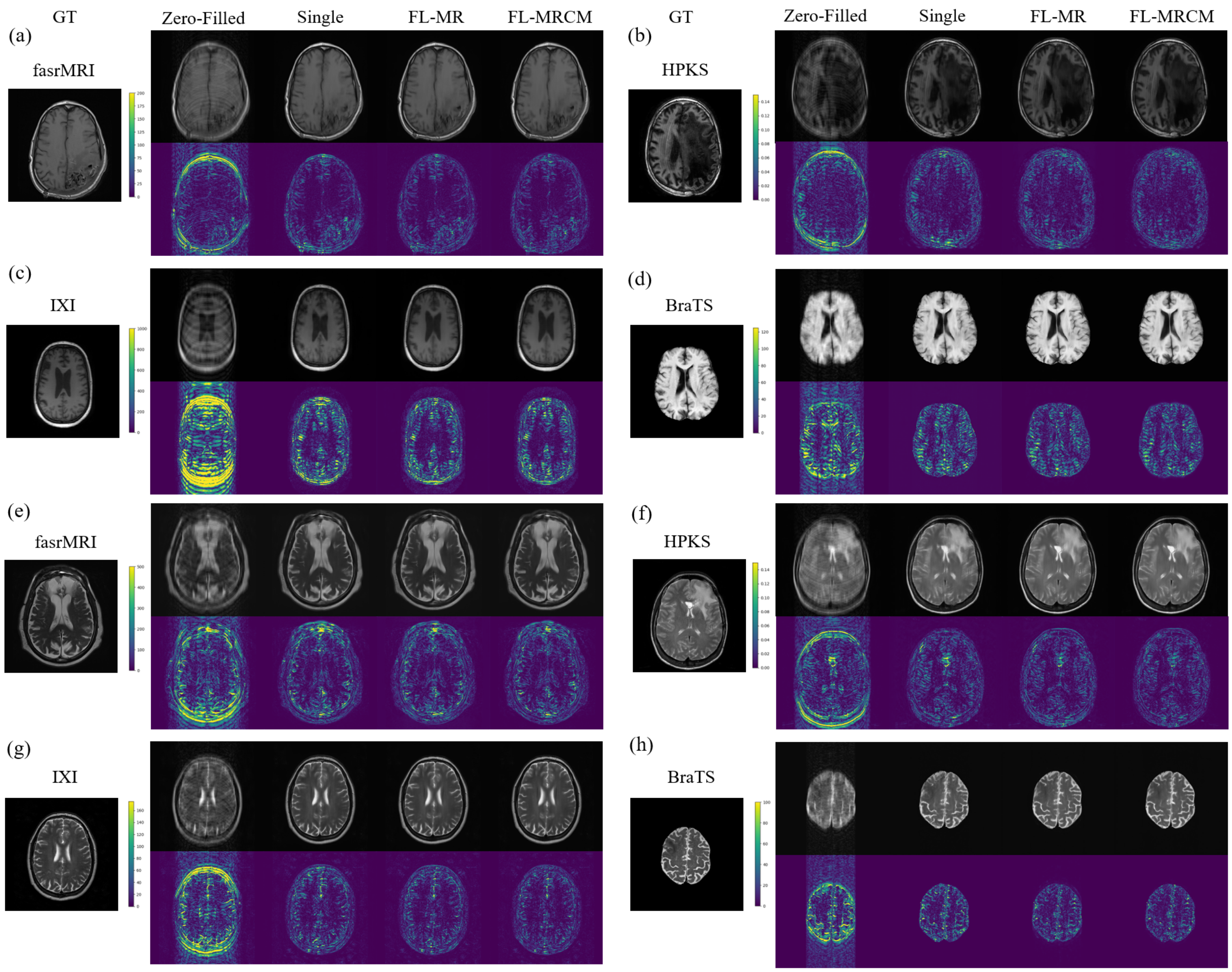

Figure 3. Qualitative results of different methods that correspond to Scenario 2. For results of T_1-weighted images on, (a) fastMRI [2], (b) HPKS, (c) IXI [1], (d) BraTS [3]. For results of T_2-weighted images on, (e) fastMRI [2], (f) HPKS, (g) IXI [1], (h) BraTS [3]. The second row of each sub-figure shows the absolute image difference between reconstructed images and the ground truth.

Table 4. Supplementary quantitative comparisons with model trained by different strategies in Scenario 1. Here, acceleration factor is set to 8.

Methods	Data Centers (Institutions)		T_1-weighted				T_2-weighted			
	Train	Test	SSIM	PSNR	Average SSIM	Average PSNR	SSIM	PSNR	Average SSIM	Average PSNR
Cross	F	B	0.8920	30.62	0.7583	25.72	0.8716	29.05	0.7583	25.72
	H	B	0.7242	27.88			0.7319	27.18		
	I	B	0.8845	29.75			0.8089	27.30		
	B	F	0.6897	23.97			0.7162	23.76		
	H	F	0.8258	28.17			0.7948	25.42		
	I	F	0.7767	26.38			0.8057	26.37		
	B	H	0.3890	21.00			0.5280	22.84		
	F	H	0.7633	24.70			0.7966	26.33		
	I	H	0.5091	22.64			0.7837	26.01		
	B	I	0.8122	26.23			0.6802	23.75		
	F	I	0.8548	27.60			0.8155	25.64		
	H	I	0.8029	26.66			0.7668	25.01		
Fused	F, H, I	B	0.8631	30.71	0.7701	27.41	0.8323	28.88	0.7966	26.76
	B, H, I	F	0.8000	27.51			0.8214	26.25		
	B, F, I	H	0.5607	23.74			0.7486	26.35		
	B, F, H	I	0.8564	27.68			0.7840	25.58		
FL-MR	F, H, I	B	0.9005	31.22	0.8339	28.08	0.8794	29.47	0.8380	27.48
	B, H, I	F	0.8598	29.14			0.8517	26.95		
	B, F, I	H	0.7178	24.15			0.7965	27.13		
	B, F, H	I	0.8574	27.80			0.8243	26.37		
FL-MRCM	F, H, I	B	0.9131	31.65	**0.8473**	**28.28**	0.8868	29.51	**0.8479**	**27.57**
	B, H, I	F	0.8697	28.75			0.8579	27.15		
	B, F, I	H	0.7440	24.72			0.8145	27.18		
	B, F, H	I	0.8625	28.01			0.8325	26.43		
Mix (Upper Bound)	F, H, I	B	0.9181	31.70	0.8545	28.45	0.8866	29.30	0.8513	27.45
	B, H, I	F	0.8690	29.33			0.8578	26.84		
	B, F, I	H	0.7726	24.94			0.8265	27.25		
	B, F, H	I	0.8581	27.85			0.8345	26.39		

Table 5. Supplementary quantitative comparisons with models trained by different strategies in Scenario 2.

Methods	Data Centers (Institutions)		T_1-weighted				T_2-weighted			
	Tain	Test	SSIM	PSRN	Average SSIM	Average PSNR	SSIM	PSNR	Average SSIM	Average PSNR
Single	B	B	0.9660	37.30	0.9351	33.81	0.9558	34.90	0.9278	32.35
	F	F	0.9494	35.45			0.9404	32.43		
	H	H	0.8855	29.67			0.9001	31.29		
	I	I	0.9396	32.80			0.9151	30.79		
Transfer	B, F	F	0.9453	34.97	0.9355	33.72	0.9353	32.03	0.9310	32.41
	B, H	H	0.8861	29.76			0.9007	31.16		
	B, I	I	0.9404	32.79			0.9119	30.65		
	F, B	B	0.9669	37.33			0.9635	35.35		
	F, H	H	0.8948	30.05			0.9153	32.10		
	F, I	I	0.9408	32.85			0.9226	31.26		
	H, B	B	0.9638	36.59			0.9566	34.47		
	H, F	F	0.9445	34.99			0.9355	31.94		
	H, I	I	0.9385	32.60			0.9162	30.63		
	I, B	B	0.9663	37.21			0.9602	34.94		
	I, F	F	0.9502	35.63			0.9376	32.38		
	I, H	H	0.8886	29.88			0.9171	32.01		
FL-MR	B, F, H, I	B	0.9662	37.37	0.9294	33.92	0.9482	35.34	0.9238	32.64
		F	0.9404	35.25			0.9306	32.19		
		H	0.8732	30.03			0.9021	31.74		
		I	0.9379	33.03			0.9145	31.29		
FL-MRCM	B, F, H, I	B	0.9676	37.57	**0.9381**	**34.14**	0.9630	35.85	**0.9373**	**33.13**
		F	0.9475	35.57			0.9385	32.69		
		H	0.8940	30.27			0.9232	32.44		
		I	0.9432	33.13			0.9244	31.54		
Mix (Upper Bound)	B, F, H, I	B	0.9698	37.62	0.9440	34.35	0.9655	35.83	0.9398	33.14
		F	0.9558	36.15			0.9435	32.82		
		H	0.9047	30.57			0.9236	32.47		
		I	0.9454	33.08			0.9266	31.44		