Multi-Modal MRI Reconstruction Assisted with Spatial Alignment Network (Supplementary Materials)

Kai Xuan, Lei Xiang, Xiaoqian Huang, Lichi Zhang, Shu Liao, Dinggang Shen, and Qian Wang

This document provides additional experiments and details for our manuscript. In Section SI, the architectures of neural networks used in our proposed method are presented in detail.

SI. Network Architectures

This section includes the designs of the neural networks used in our proposed method. The architectures of (a) the spatial alignment network $T_{\boldsymbol{\omega}}$, (b) the cross-modality synthesis network $G_{\boldsymbol{\rho}}$, and (c) the discriminator $D_{\boldsymbol{\gamma}}$ are illustrated is Fig. S1.

- The spatial alignment network, is essentially a U-Net [1] with residual connection [2], batch normalization [3], and leaky rectified linear unit activation (slope $= 0.01$) layers. Also, average pooling and nearest-neighbor upsampling are adopted to change the spatial size of feature maps. This network estimates the deformation filed ω from a pair of fully-sampled reference modality $\boldsymbol{x}_{\text{ref}}$ and under-sampled target modality images.
- The cross-modality synthesis generator is also a U-Net [1] variant with some modern tricks such as pre-activation [2] and spectral normalization [4]. It generates corresponding synthesized target modality image $\boldsymbol{x}^{\text{S}}_{\text{ref}}$ from a single reference image $\boldsymbol{x}_{\text{ref}}$. Different from $T_{\boldsymbol{\omega}}$, the generator uses convolutional layers with a stride of 2 ($s = 2$) and a kernel size of 2 ($k = 2$) instead of the average pooling layers.
- The discriminator distinguishes the synthesized target modality images $\boldsymbol{x}^{\text{S}}_{\text{tgt}}$ from those real ones $\boldsymbol{x}_{\text{ref}}$. It contains a stack of normalization-activation-convolution paradigms for feature extraction and average pooling layer (Pooling) to reduce size of feature maps. Moreover, just like the PatchGAN [5], it outputs a confidence map to the corresponding image patches. The hinge loss is used following Miyato *et al.* [4].

The competing end-to-end variational network (E2E-VarNet) [6] is used as our reconstruction backbone. Our implementation is based on the official code release[1], with minimal modification to account for multi-modal reconstruction. The network architecture is illustrated in Fig. S2, where our modification is marked in red. More specifically, to fit the multi-modal reconstruction setting, the information from the reference modality is integrated to E2E-VarNet. With the target and reference images concatenated channel-wisely, a U-net extracts complementary information from the reference modality towards better target reconstruction.

[1] https://github.com/facebookresearch/fastMRI

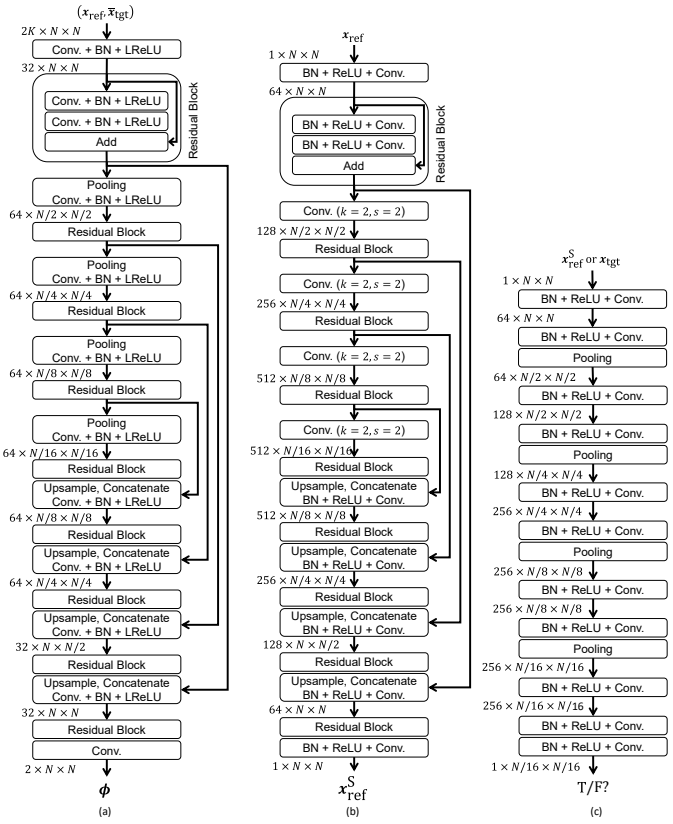

Fig. S1. Network architectures for (a) spatial alignment network $T_{\boldsymbol{\omega}}$, (b) cross-modality synthesis network $G_{\boldsymbol{\rho}}$, and (c) discriminator $D_{\boldsymbol{\gamma}}$. (a) The spatial alignment network is a U-Net variant with residual connection, batch normalization (BN), and leaky rectified linear unit (LReLU) activation (slope $= 0.01$). Also, the shapes of feature maps are explicitly specified if modified by average pooling (Pooling) or nearest-neighbor upsampling (Upsample) layers. (b) The cross-modality synthesis generator is also a U-Net variant empowered with some modern tricks. Different from $T_{\boldsymbol{\omega}}$, the generator uses convolutional layers with a stride of 2 ($s = 2$) and a kernel size of 2 ($k = 2$) instead of the average pooling layers. (c) The discriminator is relatively simple, and it contains a stack of normalization-activation-convolution paradigms for feature extraction and average pooling layer (Pooling) to reduce size of feature maps.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

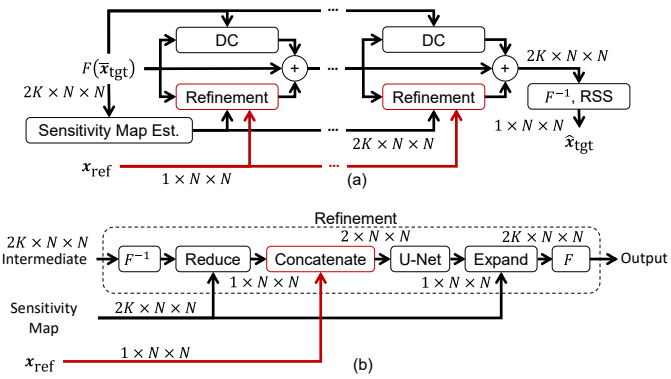

Fig. S2. Network architectures for (a) the reconstruction network $R_{\boldsymbol{\theta}}$ and (b) the zoomed-in view of the refinement module. This architecture is borrowed from E2E-VarNet [6] with minimal modifications marked in red. More specifically, to help the reconstruction of the target modality, the refinement network concatenates the target and the reference in the channel dimension before feeding a U-net.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[4] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," *InInternational Conference on Learning Representations*, Feb. 2018.

[5] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[6] A. Sriram, J. Zbontar, T. Murrell, A. Defazio, C. L. Zitnick, N. Yakubova, F. Knoll, and P. Johnson, "End-to-End Variational Networks for Accelerated MRI Reconstruction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Springer International Publishing, 2020, pp. 64–73.