(Supplementary Materials)

Improving Pelvic MR-CT Image Alignment with Self-supervised Reference-Augmented Pseudo-CT Generation Framework

Daniel Kim[1*] Mohammed A. Al-masni[2*] Jaehun Lee[3,4] Dong-Hyun Kim[1†] Kanghyun Ryu[4†]
[1]Yonsei University [2]Sejong University [3]Yale University [4]Korea Institute of Science and Technology
[1]{danny4159, donghyunkim}@yonsei.ac.kr, [2]m.almasani@sejong.ac.kr,
[3]jaehun.lee@yale.edu, [4]khryu@kist.re.kr

1. Details on Synthesis network in Stage 1

1.1. Brief Explanation

In this section, we will provide a more detailed explanation of the learning scheme for the synthesis network in Stage 1. The notations follow those defined in the main manuscript. The primary objective of this network is to effectively disentangle the structural (content) information and contrast (style) information from both MR and CT images. By integrating the structural information from MR with the contrast information from CT, we aim to generate an initial synthetic CT (synth-CT) that is structurally coherent with MR. Although the CT style is applied globally, the generated synth-CT may lack accurate textures and sharp details. These limitations can be addressed in Stage 2 through feature aggregation using the DACA block.

Since the architecture of the network has already been described in the main manuscript, we will focus here on explaining the three composite loss functions used to train the network.

In this network, the encoders and decoders consist of convolutional modules. The style encoder applies global average pooling at the final layer to extract global style features, which capture the contrast or intensity information from the CT image. The encoders follow the same notation as in Figs. 1, 2, and 3, and share parameters across these figures.

The losses are composed of self-reconstruction loss, latent-reconstruction loss, and cycle consistency loss. The purpose of self-reconstruction loss is to enhance the feature extraction capability of the encoder through identity reconstruction for each image. This will prevent the encoder from losing important features. The latent reconstruction loss aims to enhance disentangled representation by regularizing the content and style codes to maintain consistency in the latent space. For the content code, the code (c_{MR}) extracted from I_{MR} and the code ($c'CT$) extracted from the image ($\acute{I}CT$), which is generated using the same c_{MR}, should be identical in the latent space. For the style code, if the input data and the generated data belong to the same domain, their style codes should also be identical in the latent space. Cycle consistency loss Since the data pairs are misaligned, perfect ground truth is unavailable, so we applied the widely-used cycle consistency loss in unsupervised learning. As shown in Fig. 3, we used s_{MR} instead of s'_{MR} when reconstructing $\tilde{I}CT$ to reduce training complexity, while the latent consistency loss ensured both elements were identical.

1.2. Self reconstruction loss

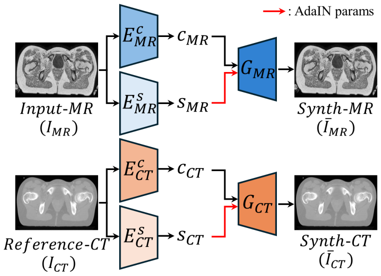

Figure 1. Self reconstruction loss for training the synthesis network

$$\begin{aligned}\mathcal{L}^{I_{MR}}_{\text{recon}} &= \mathbb{E}\left[\|G_{MR}(c_{MR}, s_{MR}) - I_{MR}\|_1\right] \\ &= \mathbb{E}\left[\left\|\bar{I}_{MR} - I_{MR}\right\|_1\right]\end{aligned}$$

$$\begin{aligned}\mathcal{L}^{I_{CT}}_{\text{recon}} &= \mathbb{E}\left[\|G_{CT}(c_{CT}, s_{CT}) - I_{CT}\|_1\right] \\ &= \mathbb{E}\left[\left\|\bar{I}_{CT} - I_{CT}\right\|_1\right]\end{aligned}$$

[*] Equal contribution
[†] Co-corresponding author

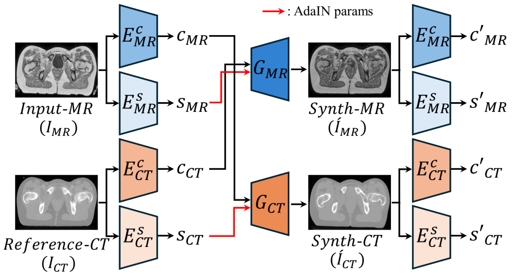

Figure 2. Latent reconstruction loss for training the synthesis network

1.3. Latent reconstruction loss

$$\begin{aligned}\mathcal{L}^{c_{MR}}_{\text{recon}} &= \mathbb{E}\left[\|E^c_{CT}(G_{CT}(c_{MR}, s_{CT})) - c_{MR}\|_1\right] \\ &= \mathbb{E}\left[\|c'_{CT} - c_{MR}\|_1\right]\end{aligned}$$

$$\begin{aligned}\mathcal{L}^{c_{CT}}_{\text{recon}} &= \mathbb{E}\left[\|E^c_{MR}(G_{MR}(c_{CT}, s_{MR})) - c_{CT}\|_1\right] \\ &= \mathbb{E}\left[\|c'_{MR} - c_{CT}\|_1\right]\end{aligned}$$

$$\begin{aligned}\mathcal{L}^{s_{MR}}_{\text{recon}} &= \mathbb{E}\left[\|E^s_{MR}(G_{MR}(c_{CT}, s_{MR})) - s_{MR}\|_1\right] \\ &= \mathbb{E}\left[\|s'_{MR} - s_{MR}\|_1\right]\end{aligned}$$

$$\begin{aligned}\mathcal{L}^{s_{CT}}_{\text{recon}} &= \mathbb{E}\left[\|E^s_{CT}(G_{CT}(c_{MR}, s_{CT})) - s_{CT}\|_1\right] \\ &= \mathbb{E}\left[\|s'_{CT} - s_{CT}\|_1\right]\end{aligned}$$

1.4. Cycle consistency loss

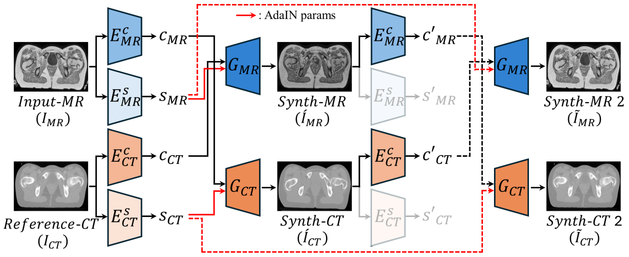

Figure 3. Cycle consistency loss for training the synthesis network

$$\tilde{I}_{MR} = G_{MR}\Big(E^c_{CT}\big(G_{CT}\big(E^c_{MR}(I_{MR}), s_{CT}\big)\big), E^s_{MR}(I_{MR})\Big)$$

$$\mathcal{L}^{I_{MR}}_{\text{cycle}} = \mathbb{E}\left[\left\|\tilde{I}_{MR} - I_{MR}\right\|_1\right]$$

$$\tilde{I}_{CT} = G_{CT}\Big(E^c_{MR}\big(G_{MR}\big(E^c_{CT}(I_{CT}), s_{MR}\big)\big), E^s_{CT}(I_{CT})\Big)$$

$$\mathcal{L}^{I_{CT}}_{\text{cycle}} = \mathbb{E}\left[\left\|\tilde{I}_{CT} - I_{CT}\right\|_1\right]$$

2. Underestimated Segmentation Accuracy

The Dice coefficients presented in Table 2 of the main manuscript are generally lower due to inaccuracies in the ground truth. Although MR images offer superior soft tissue contrast, they produce weak signals from bones, which can result in errors when using TotalSegmentator to obtain segmentation labels. As shown in Figure 4, the masking results of the generated pseudo-CT are precise, but errors in the ground truth result in underestimated segmentation accuracy.

3. Metric: Gradient correlation(GC)

To evaluate the structural correlation between different imaging modalities, we used Gradient Correlation (GC) with Canny edge detection. We set the threshold for MR images to (170, 190) to capture strong edges, CT images to (30, 50). The visualizations are shown in Fig. 5.

4. Additional Comparison with Registration Methods

In MR images, bone structures are often not clearly distinguishable from surrounding tissues, which can lead to errors in bone registration. However, our proposed method successfully aligns the bone areas with the estimated bone regions in the MRI, as demonstrated in the zoomed view in Fig. 6. Further results are presented in Figs. 7 and 8.

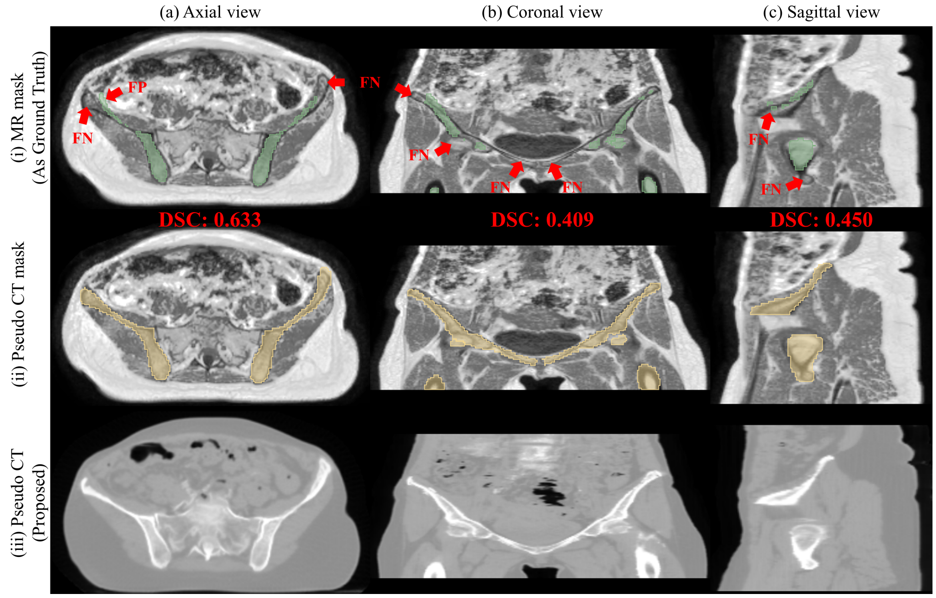

Figure 4. Visualization of segmentation masks for the bones (hips, femurs). (i) Ground truth segmentation mask obtained from MR is overlaid on the MR image. (ii) Segmentation mask obtained from (iii) is overlaid on the MR image.

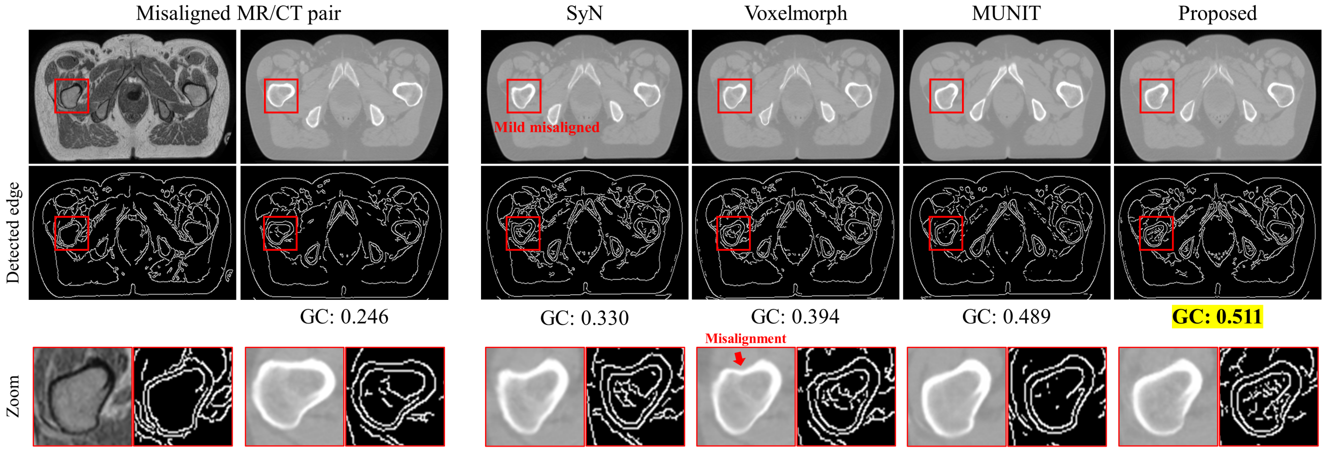

Figure 5. Visualization of Gradient Correlation Metric(GC)

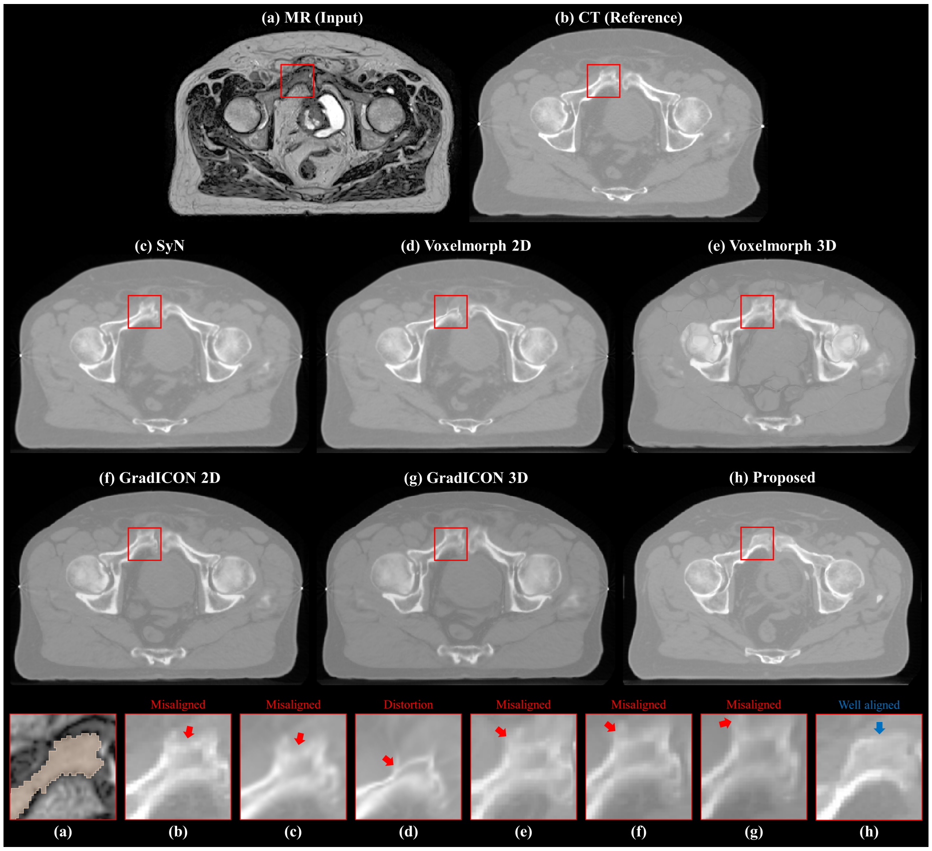

Figure 6. Comparison of Registration methods (SyN, Voxelmorph, GradICON) against ours.

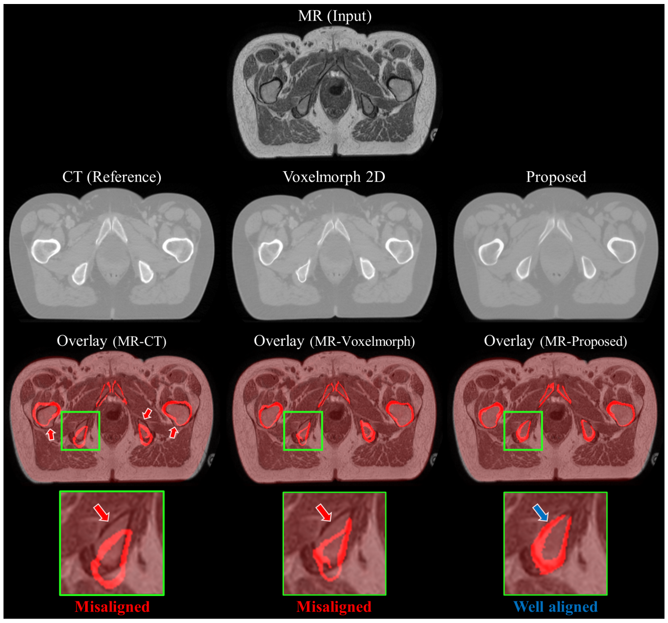

Figure 7. Comparison of reference CT, registration method (VoxelMorph), and ours. CT values are thresholded to highlight bone regions in red and overlaid onto MR images.

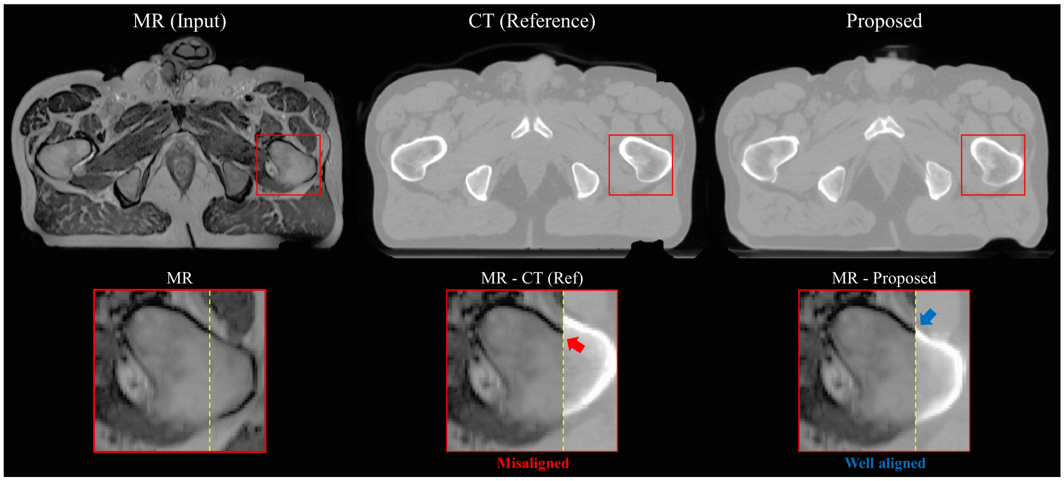

Figure 8. Comparison of reference CT and ours, highlighting bone region misalignment.