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

2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3315-1083-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/WACV61041.2025.00044

Abstract

RegistFormer, our novel reference-augmented image synthesis framework, generates aligned pseudo-CT images (with respect to MR) from misaligned MR and CT pairs. RegistFormer addresses the limitations of intensity-based registration methods, which often fail due to dissimilar image features and complex deformation fields. Unlike conventional image-to-image (I2I) translation methods, our method uses a misaligned CT scan as an auxiliary input to guide the synthesis task through the Deformation-Aware Cross-Attention (DACA) mechanism. DACA integrates the deformation field from a registration method to aggregate spatially matched features from the misaligned CT into MR spatial coordinates. Additionally, we propose a novel combination of loss functions for training with datasets of misaligned MR-CT pairs in a self-supervised manner, eliminating the need for pre-aligned training data. Experiments were conducted with the synthRAD2023[1] MR-CT pelvis pair dataset. RegistFormer outperforms past state-of-the-art methods, including I2I, registration, and hybrid (registration + I2I), across metrics evaluating both structure alignment and distribution similarity. Moreover, RegistFormer demonstrates superior performance in zero-shot segmentation downstream tasks, highlighting its clinical value. Source code: https://github.com/danny4159/RegistFormer

1. Introduction

In general, registration is an important preprocessing step in medical image processing because it integrates useful information from two or more images into a single, more informative image. Registration of pelvic CT and MRI scans is useful as it facilitates the effective fusion of these two modalities for prostate cancer radiation therapy [12], using CT for dose planning and MRI for precise organ delineation. However, this registration is challenging due to inter-modality appearance differences and shape/appearance variations of pelvic organs [12]. Even for the same patient, CT and MRI scans, acquired using different devices, can exhibit substantial changes in organ shapes due to inevitable rectal movement and varying bladder states. The spatial proximity of the prostate, bladder, and rectum means that a shape change in one organ often impacts the neighboring organs.

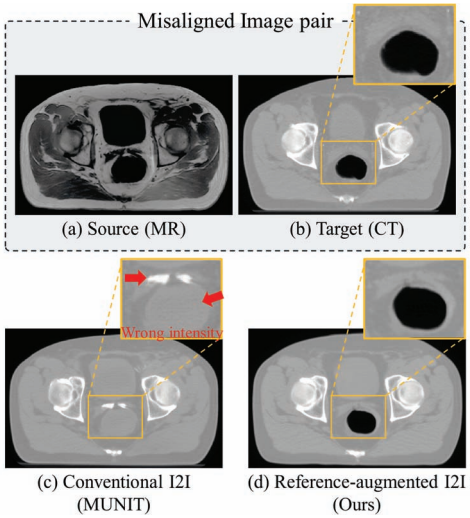

Figure 1. Illustrative comparison highlighting limitations of conventional I2I translation methods in contrast to our approach.

Non-rigid registration methods like SyN [4] and NiftyReg [5] can estimate local deformations due to tissue movement, but this is challenging due to the high dimensionality and non-linear relationship between image contents. Large anatomical deformations, different patient positions, and varying hollow organ fillings, particularly affecting prostate position, further complicate the task. Sup-

*Equal contribution
†Co-corresponding author
[1]https://synthrad2023.grand-challenge.org/

2642-9381/25/$31.00 ©2025 IEEE
DOI 10.1109/WACV61041.2025.00044

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

plementary methods like manual delineation [7] or intraprostatic fiducial markers [3] add effort and have limitations. Recently, learning-based techniques using various networks [19, 32, 37] have been introduced, offering speed advantages. However, their accuracy often does not significantly surpass classical optimization-based approaches [34].

I2I translation frameworks can generate pseudo-CT images to replace registered CT images or convert a multi-modality problem into a uni-modality problem for registration tasks [30]. However, these approaches, including MR to CT synthesis, often suffer from issues like blurriness [21] and hallucinations [27]. The most significant risk is hallucination, where non-existent content is generated or relevant information is removed. Misaligned I2I translations, often using cycle-consistent GANs trained with unsupervised data, are particularly susceptible due to the lack of direct verification that the synthesized image accurately represents the original content.

Our method RegistFormer, is a novel approach to enhancing the I2I framework by incorporating an auxiliary misaligned CT image, termed reference-augmented pseudo-CT generation. This addition improves robustness and reduces erroneous outputs. Fig. 1(c) shows errors from the original CT synthesis method, such as distinguishing hollow organ fillings from soft tissue or holes in MR images, which are resolved more effectively using CT images.

Our method designs a cross-attention-based transformer module to transfer intricate semantic features from reference CT images to the MR coordinate system. To address MR and CT coordinate discrepancies, we use the deformation field estimated by a registration network, applying windowed cross-attention at matching locations. This module operates within the multi-scale semantic feature space of the U-Net architecture, performing a "copy and paste" operation across multiple scales.

Training the framework in a self-supervised manner is required due to the rarity of perfectly aligned data pairs in in-vivo subjects. Thus, we design a method that can use the misaligned reference CT images as the ground truth by the combination of loss functions. We propose a combination of a structure-preserving loss derived from contrastive loss [23], a misalignment-robust loss function [16] to be tolerant to misalignment, and an adversarial loss to improve image quality.

In summary, the main contributions of this paper are:

1. We design a novel deformation-aware attention module that enhances the windowed cross-attention mechanism by transferring spatially matched features, resulting in improved output.

2. We propose a novel combination of loss functions tailored for self-supervised learning with misaligned MR-CT pairs. This includes structure-preserving and misalignment-robust feature loss functions, which ensure that the network generates well-aligned images without the need for pre-aligned training data.

3. We demonstrate the comprehensive superiority of our method through rigorous evaluation across diverse metrics for structural alignment and texture preservation.

2. Related work

2.1. Methods for aligning image pairs

Image-to-image(I2I) translation: Unsupervised I2I translation learns the conditional translation of images between domains without relying on explicitly paired datasets. This is suitable for tasks where data pairs are misaligned. In the medical domain, I2I translation is often known as image synthesis. One popular approach is CycleGAN [11], which uses cycle consistency loss to train a network that preserves structural features while translating images between domains. To further improve structural preservation during translation, subsequent studies such as GCGAN [13], SC-GAN [25], and CUT [23] have introduced additional constraints. More recent methods based on representation disentanglement, such as UNIT [9], MUNIT [14], and DRIT [15], decompose images into content and style codes, which are then recombined to generate new images. However, all these methods operate under the assumption of an evident one-to-one mapping between modalities. This assumption is violated in the context of CT and MR due to their fundamentally distinct physical principles.

Registration method: Traditional optimization-based methods calculate deformation fields on the fly, with notable examples including the elastic-type model [2], B-spline-based model [1], and Standard Symmetric Normalization (SyN) [4]. On the other hand, learning-based methods leverage deep learning networks to generate deformation fields, with key approaches like VoxelMorph [18] and its extensions, including Dual-PRNet [35], LapIRN [22] and TransMorph [32]. Recently, Grad-ICON [40] has further enhanced stability by regularizing the deformation field using gradient inverse consistency. However, MR to CT alignment, particularly in the pelvic region, remains challenging due to significant deformations and contrast differences, making the task inherently ill-posed.

Hybrid method: Hybrid methods that integrate cross-modality I2I translation and registration into a single architecture have recently become prominent [28, 36, 39]. For MR to CT tasks, I2I translation can first convert MR images to the CT modality before registration [33, 26]. However, even with accurate synthesized CT generation, estimating a dense deformation field is challenging due to its high dimensionality and the complex, non-linear relationship between synthesized and original images. Furthermore, errors

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

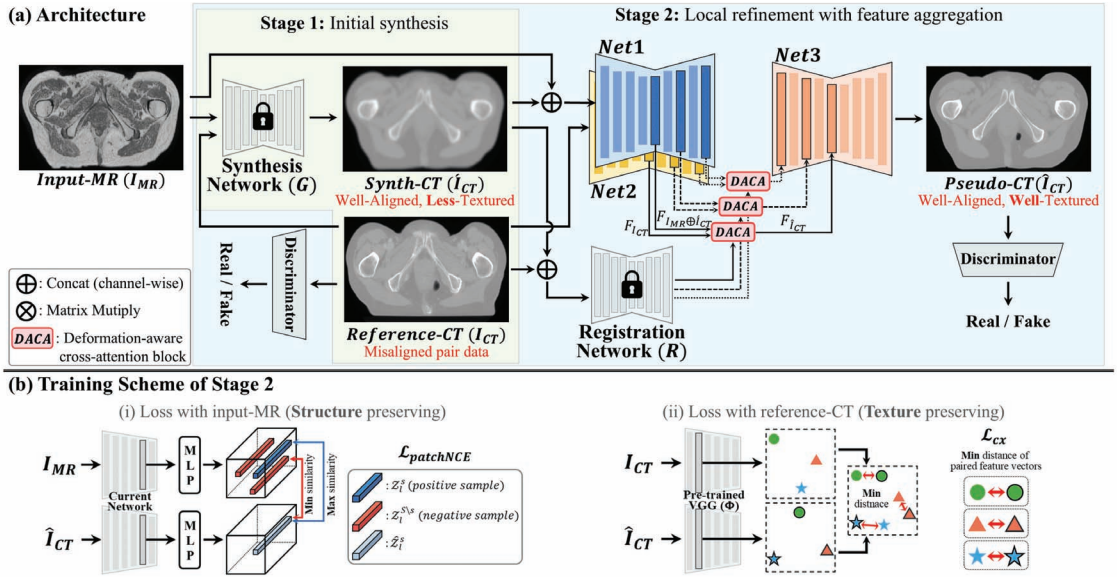

Figure 2. (a) illustrates the overall architecture of the proposed framework, where synth-CT is generated in Stage 1, followed by feature aggregation from the reference-CT in Stage 2 through the DACA block to produce the final pseudo-CT. (b) visualizes the operation of the loss functions used for training Stage 2. The combination of the two losses, illustrated in (b-i) and (b-ii), aims to generate a pseudo-CT that is aligned with MR while preserving the texture of CT.

in the I2I translation can propagate to the subsequent registration process, compromising overall alignment accuracy.

2.2. Reference-based super resolution

Our research is primarily inspired by the study from reference-based super-resolution (RefSR) literature, which enhances image resolution quality using reference images related to the low-resolution input. These approaches, designed for super-resolution of natural images, as well as for image restoration, leverage features from high-quality reference images to enhance the quality of the target images [24, 31, 38].

While we draw inspiration from these methods, our approach introduces a novel application in the different domain of medical imaging, specifically for MR to CT alignment task. Unlike traditional RefSR techniques that focus solely on enhancing image resolution, the primary focus of our method is to generate a pseudo-CT that is aligned with MR.

3. Method

3.1. Two-stage framework

The proposed framework, illustrated in Fig. 2(a), consists of two primary stages. The notation used in Figs. 2, 3, and 4 is consistent with the equations presented in this section.

The first stage employs a synthesis network, leveraging the fact that registering unimodal images is simpler than registering multimodal images [33]. In this stage, the synthesis network generates an initial synth-CT ($\acute{I}_{CT}$) image from the input-MR (I_{MR}) image. Unlike conventional methods, our approach incorporates a reference-CT (I_{CT}) image to provide global style information, thereby mitigating the risk of generating erroneous outputs.

In the second stage, the generated synth-CT images are refined using the reference-CT images through a Deformation-Aware Cross-Attention (DACA) block. This refinement occurs locally within a multi-resolution feature space, enhancing the accuracy of the synthesized image. The DACA block applies windowed attention [29] to corresponding regions, guided by the deformation field (ϕ) provided by the registration network (R). The deformation field (ϕ) ensures that the attention mechanism aligns features accurately across images. Let $F_{I_{MR}\oplus\acute{I}CT}$ represent the feature map of the concatenated dual channels of input-MR and synth-CT, and $F_{I_{CT}}$ represent the feature map of the reference-CT. The refined feature map $F_{\hat{I}_{CT}}$ is obtained as:

$$F_{\hat{I}_{CT}} = DACA(F_{I_{MR}\oplus\acute{I}_{CT}}, F_{I_{CT}}, \phi) \quad (1)$$

The final pseudo-CT ($\hat{I}_{CT}$) image is then reconstructed from the refined feature map by :

$$\hat{I}_{CT} = Net3(F_{\hat{I}_{CT}}) \quad (2)$$

This two-stage framework, incorporating both global style information and local feature refinement, can enhance the robustness and accuracy of MR to CT image synthesis and registration.

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

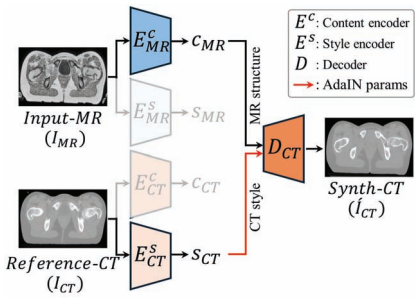

Figure 3. Illustration of the synthesis network in Stage 1 for generating the initial synth-CT ($\acute{I}_{CT}$).

Stage 1: Initial synthesis using global style transfer

The primary objective of the synthesis network (G) is to generate an initial synth-CT ($\acute{I}_{CT}$) image aligned with the input-MR (I_{MR}) image (see Fig. 2(a)). Therefore, the synthesis network operates to preserve the structural information from the input-MR (I_{MR}) images while incorporating style information from the reference-CT (I_{CT}) images.

Inspired by the concept of disentangled representation introduced in MUNIT [14], we designed the synthesis network to separate structural information (i.e., content code) and contrast information (i.e., style code) from each image through two encoders. As illustrated in Fig. 3, the content code from the MR image (c_{MR}) and the style code from the CT image (s_{CT}) are combined through the decoder (D) to generate the synth-CT ($\acute{I}_{CT}$) image. The style code globally transfers style information using the adaptive instance normalization (AdaIN) method [8] during the decoding process, as follows:

$$\text{AdaIN}(c_{\text{MR}}, s_{\text{CT}}) = \sigma(s_{\text{CT}}) \cdot \left(\frac{c_{\text{MR}} - \mu(c_{\text{MR}})}{\sigma(c_{\text{MR}})} \right) + \mu(s_{\text{CT}}) \tag{3}$$

The learning scheme of this synthesis network is described in detail in Sec. 1 of the Supplementary Material (Supp.).

This generated synth-CT ($\acute{I}_{CT}$) image is then utilized in subsequent processing stages to refine and improve the final output. This approach ensures that the initial synthesis maintains the structural alignment with the input-MR (I_{MR}) image while incorporating the global style characteristics of the reference-CT (I_{CT}) image, thereby addressing issues of incorrect intensity and hallucination.

Stage 2: Local refinement with feature aggregation

The purpose of this stage is to refine the initial synth-CT ($\acute{I}_{CT}$) image obtained from the synthesis network (G) by aggregating features from the neighborhood of the "corresponding pixel" in the reference-CT (I_{CT}) image. The corresponding pixel refers to the redefined coordinates derived from the deformation field (ϕ).

First, the deformation field (ϕ) is obtained from a pre-

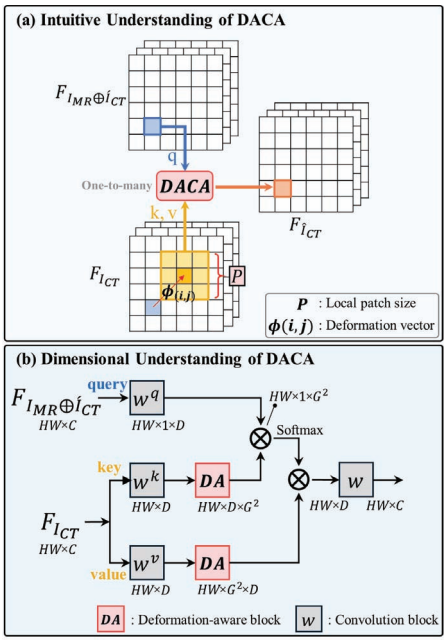

Figure 4. The DACA block for aligned feature aggregation of misaligned image pairs. (a) illustrates the operation of the DACA block from a pixel perspective. (b) shows the DACA block from a dimensional perspective, implemented by applying the DA block to the cross-attention mechanism.

trained registration network (R) between the moving image (synth-CT) and the fixed image (reference-CT). During Stage 2 training, the registration network (R) remains frozen and is used solely to infer the deformation field (ϕ) for subsequent steps. This is a key difference from the registration approach, which optimizes the deformation field through training. For faster inference, we recommend using learning-based approaches like Voxelmorph [19], which we employed. The deformation field (ϕ) consists of two channels, representing displacements in the x and y directions, respectively:

$$\begin{aligned} \phi(i,j) &= R(\acute{I}_{CT}, I_{CT}) \\ (i',j') &= (i + \phi_w(i,j), j + \phi_h(i,j)) \end{aligned} \tag{4}$$

Next, two U-Nets ($Net1$, $Net2$) are used for multi-scale feature extraction. In $Net1$, the input-MR and synth-CT are processed as a dual-channel input, extracting the feature map $F_{I_{MR} \oplus \acute{I}CT}$, while $Net2$ processes the reference-CT as a single-channel input, yielding the feature map $F_{I_{CT}}$. Following the standard transformer [10], which enhances performance through various representations, $F_{I_{MR} \oplus \acute{I}CT}$ is treated as the "query" using w^q, while $F_{I_{CT}}$ serves as the "key" and "value", with distinct learnable parameters w^k and w^v (see Fig. 4(b)). The query is based on the original coordinates, while the key and value use the coordinates redefined by the deformation field (refer to Eq. (4)).

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

Table 1. Quantitative comparison of various methods. Blue denotes the top-performing method, while green indicates the second-best.

Types	Methods	Gradient Correlation ↑ (with MR)	Normalized Mutual Information ↑ (with MR)	FID ↓ (with ref-CT)	Sharpness ↑ (no reference)
Original	Ref-CT	0.23 ± 0.05	0.31 ± 0.54	–	87.1 ± 41.5
Unsupervised I2I	CGAN [20]	0.15 ± 0.05	0.33 ± 0.03	60.6	66.2 ± 19.5
	SCGAN [25]	0.11 ± 0.04	0.29 ± 0.03	123.9	79.5 ± 28.7
	UNIT [9]	0.45 ± 0.06	0.36 ± 0.03	32.2	74.2 ± 24.8
	MUNIT [14]	0.34 ± 0.06	0.33 ± 0.03	28.9	80.6 ± 24.3
Registration	SyN [4]	0.28 ± 0.05	0.45 ± 0.05	9.9	55.0 ± 18.4
	Voxelmorph(2D) [18]	0.30 ± 0.06	0.37 ± 0.06	20.2	74.3 ± 25.6
	Voxelmorph(3D) [18]	0.28 ± 0.08	0.37 ± 0.05	32.9	88.5 ± 21.5
	LapIRN(3D) [22]	0.29 ± 0.07	0.37 ± 0.05	34.7	71.95 ± 14.2
	TransMorph(3D) [32]	0.31 ± 0.08	0.38 ± 0.05	35.76	106.0 ± 20.9
	GradICON(2D) [40]	0.28 ± 0.06	0.36 ± 0.05	21.6	61.7 ± 16.2
	GradICON(3D) [40]	0.26 ± 0.06	0.36 ± 0.04	26.3	85.0 ± 19.8
Registration + I2I	RegGAN [28]	0.38 ± 0.05	0.40 ± 0.04	59.1	89.1 ± 14.6
Proposed	RegistFormer	0.43 ± 0.06	0.40 ± 0.04	19.2	89.3 ± 13.5

A one-to-many windowed cross-attention [29] is performed between each pixel in $F_{I_{MR} \oplus \acute{I}_{CT}}$ and the surrounding local patch (of size P) of the corresponding pixel in $F_{I_{CT}}$ (see Fig. 4(a)). This one-to-many operation helps to compensate for errors in the deformation field. This process, which we term the Deformation-Aware Cross-Attention (DACA) block, is as follows:

$$F_{\hat{I}_{CT}}(i,j) = \sum_i \sum_j \text{Attention}(F_{I_{MR} \oplus \acute{I}_{CT}}(i,j), F_{I_{CT}}(i'+p, j'+p)) \quad \text{for all } p \in \{-\lfloor P/2 \rfloor, \ldots, \lfloor P/2 \rfloor\}, p \in \mathbb{Z} \tag{5}$$

This feature aggregation process is applied across multiple layers to ensure that features from the reference-CT image are well reflected at various resolution levels. The refined feature map ($F_{\hat{I}_{CT}}$) is then used for the encoding part of a third U-Net ($Net3$), ultimately generating the pseudo-CT ($\hat{I}_{CT}$), as shown in Eq. (2).

3.2. Self-supervised training scheme for Stage 2

Loss with input-MR (Fig. 2(b-i)): To maintain structural alignment with the input-MR images, we used patch-based contrastive loss (patchNCE [23]) based on noise-contrastive estimation (InfoNCE [17]). InfoNCE acts as a lower bound to mutual information (MI), preventing the network from trivially outputting the reference-CT and instead generating a pseudo-CT aligned with the input-MR images. The loss function is:

$$\mathcal{L}_{\text{patchNCE}}(\hat{I}_{CT}, I_{MR}) = \mathbb{E}\left[\sum_{l=1}^{L}\sum_{s=1}^{S_l} \ell(\hat{z}_l^s, z_l^s, z_l^{S \setminus s})\right]$$

Here, ℓ maximizes the cosine similarity between the feature vector ($\hat{z}_l^s$) from the pseudo-CT ($\hat{I}_{CT}$) and the corresponding feature vector (z_l^s) from the input-MR (I_{MR}), while minimizing similarity with feature vectors ($z_l^{S \setminus s}$) from other locations in the input-MR. Feature vectors are obtained by sampling patches from the feature map, extracted through a VGG-19 network, and vectorizing them through an MLP. This mechanism ensures alignment and penalizes different locations in the feature space, thereby enhancing structural alignment between cross-modality images.

Unlike the original PatchNCE loss [23], which uses the generator's encoder for feature extraction, our method employs a pre-trained VGG-19 network instead, resulting in better performance and less computation. To prevent the direct transfer of input-MR structures, we use the "conv4_2" and "conv5_4" layers, which capture high-level features.

Loss with reference-CT (Fig. 2(b-ii)): In real-world scenarios, acquiring aligned CT images is often infeasible, resulting in reliance on misaligned data [19]. Consequently, in this context, we must utilize misaligned reference-CT images as proxy labels and train it with self-supervision.

Common loss functions such as L_1 and L_2 are unsuitable due to misalignment. Similarly, perceptual loss, which measures high-level feature differences, is more robust to minor misalignments but still struggles with significant misalignments.

Therefore, for the loss between the generated output and the reference-CT, we employed the contextual loss [16], which is resilient to misalignment. For two images $\hat{I}_{CT}$ and I_{CT}, the contextual loss $\mathcal{L}_{cx}(\hat{I}_{CT}, I_{CT})$ aims to minimize the cumulative distance of all matched feature pairs, as formulated below:

$$\mathcal{L}_{cx}(\hat{I}_{CT}, I_{CT}) = \mathbb{E}\left[-\frac{1}{N}\sum_i \min_j \psi(\Phi(\hat{I}_{CT})_i, \Phi(I_{CT})_j)\right]$$

Here, $\Phi(\cdot)$ represents the feature maps extracted from layer l of the pre-trained VGG-19 network (Φ), and ψ measures the cosine similarity loss. The `min` operation compares feature vector i of image $\hat{I}_{CT}$ with all feature vectors of image I_{CT} and pairs it with the one that has the smallest

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

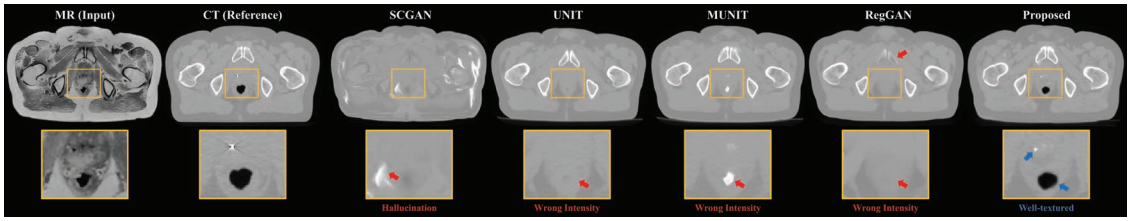

Figure 5. Comparison of I2I synthesis methods (SCGAN, UNIT, MUNIT, RegGAN) against ours

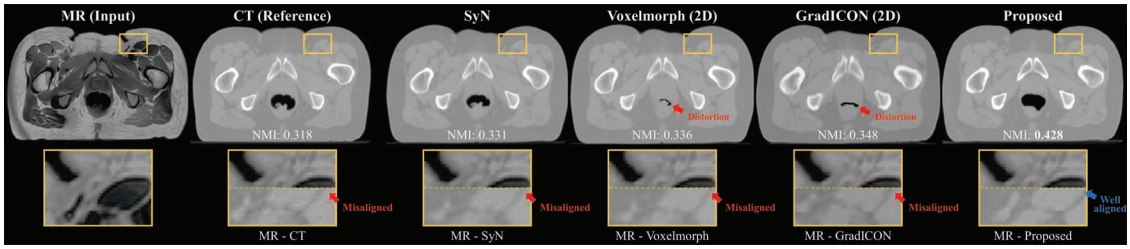

Figure 6. Comparison of Registration methods (SyN, Voxelmorph, GradICON) against ours. Normalized Mutual Information values are overlaid

distance, thereby addressing the issue of misalignment. In our experiments, we utilized multi-scale feature maps from the layers "conv1_2", "conv2_2", and "conv3_2" to capture high-resolution features, allowing for finer texture representation.

To enhance the generation of high-fidelity images, we employed a discriminator loss based on a Generative Adversarial Network (GAN). This approach encourages the generation of images that are indistinguishable by the discriminator from the reference-CT. Specifically, we utilized the patch-based Least Squares GAN (LSGAN) method [11].

Total loss function: The total loss is computed as follows, with the balance of the weights being task-dependent.

$$\mathcal{L}_{total} = \lambda_{\text{patchNCE}}\mathcal{L}_{patchNCE} + \lambda_{\text{cx}}\mathcal{L}_{cx} + \lambda_{\text{gan}}\mathcal{L}_{gan}$$

4. Experimental setup

4.1. Dataset and pre-processing

We utilized the publicly available SynthRAD2023 dataset[2] with MR to CT data pairs in the pelvis region. Despite initial rigid registration, significant misalignments remained, providing a realistic challenge. MR images were sourced from three institutions using T1-weighted gradient echo sequences at 1.5T or 3T. We pre-processed MR data with N4 bias field correction and Nyul histogram matching to standardize contrast variations. For CT images, we applied windowed HU clipping based on the 5% to 95% percentiles and performed z-score normalization. From a total of 165 patients, excluding participants with severe artifacts, 112 were used for the training set, 18 for the validation set, and 35 for the testing set. All images were resized to ensure consistent width and height across all patients. During training, we used random patch cropping with a fixed size of 96×96.

[2] https://synthrad2023.grand-challenge.org/

4.2. Implementation details

We first train the Synthesis Network (G) and freeze it. Then, we train the Registration Network (R), freeze it, and proceed to train the other untrained part of the network. For cross-attention, we use two heads with a local patch size of 28 (see Table 3). The U-Net is fully convolutional and initialized using Kaiming normal initialization. We use the Adam optimizer with default settings and a multistep learning rate scheduler to reduce the learning rate by 0.1 at the 20th and 30th epochs. Total loss weights $\lambda_{\text{patchNCE}}$, λ_{cx}, and λ_{gan} are set to 0.1, 1, and 0.1, respectively.

Experiments were conducted on NVIDIA RTX A5000 GPUs using PyTorch, with a batch size of 4 for 100 epochs, taking 68 hours on a single GPU.

5. Results

5.1. Evaluation of pseudo-CT quality

Experimental setting: We compared the image quality of pseudo-CTs generated by our method with widely used and state-of-the-art unsupervised I2I translation methods [20, 25, 9, 14], registration methods [4, 18, 22, 32, 40], and a hybrid approach [28], as shown in Table 1. Our main objective is to ensure high structural alignment with MR images while maintaining accurate texture and details. We

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

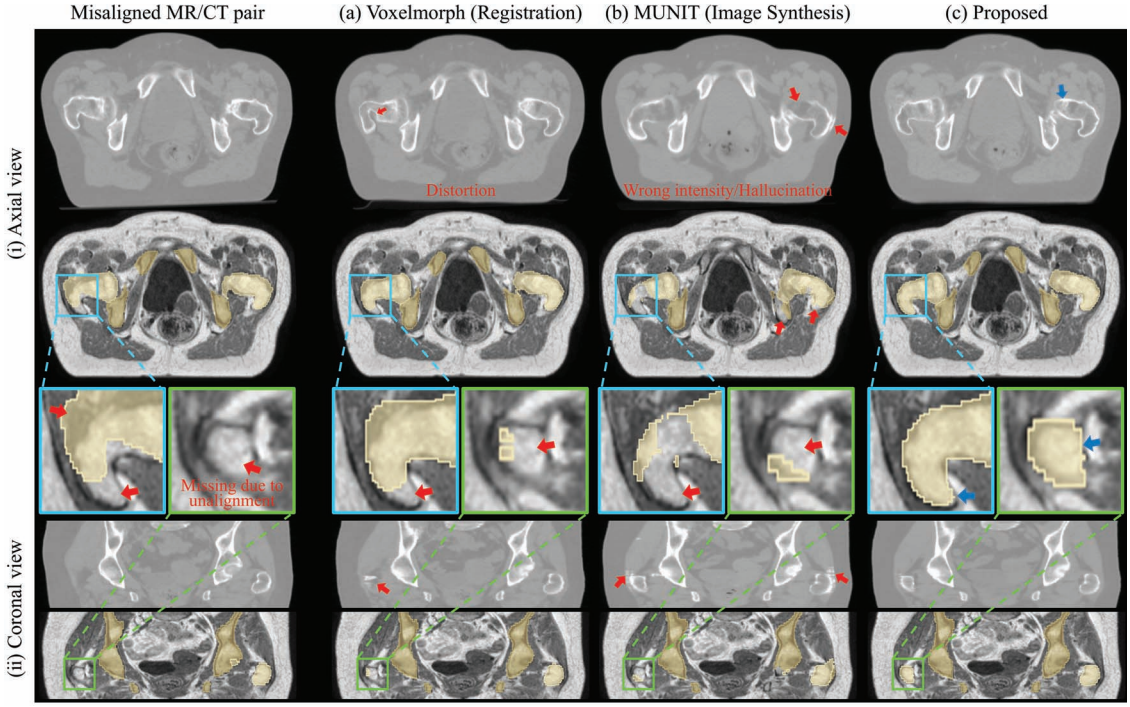

Figure 7. Comparative analysis of downstream segmentation from 3D stacked pseudo-CT image. The yellow regions indicate the bone segmentation masks produced by TotalSegmentator. 3D images are obtained by stacking 100 axial images.

assess structural alignment using gradient correlation (GC; demonstrated in Sec. 3 of the Supp) with Canny edge detection and normalized mutual information (NMI). Texture quality relative to reference CTs is measured using Fréchet Inception Distance (FID) on the test data. Image sharpness, without any reference, is evaluated by calculating the variance of the laplacian, with higher values indicating sharper images.

Results: In Table 1, our proposed method demonstrates second-best performance across all metrics, including GC, NMI, FID, and sharpness. While UNIT achieves the highest GC of 0.45, it frequently exhibits hallucinations and incorrect intensity values, which is partially indicated by its lower FID score (see Fig. 5). SyN shows the best performance in NMI and FID but suffers from significant blurring during interpolation, resulting in a low sharpness score of 55.0. Notably, despite being a 2D network, our method surpasses the performance of 3D registration networks across all metrics. Qualitative evaluations, illustrated in Figs. 5 and 6, demonstrate the effectiveness of our method in preserving high-quality textures and accurate alignment.

5.2. Segmentation comparison

Experimental setting: To demonstrate the potential applications of our method, we evaluated its effectiveness in segmentation using pseudo-CT images with TotalSegmentator [41], a tool designed for segmenting major anatomical structures in CT images. For MR image segmentation, we used a recently published version of the same tool [42]. Axial images were stacked in 3D before processing. For comparison, we performed an identical segmentation process on results obtained from the comparison models: registration and I2I translation approaches. The segmentation outcomes were qualitatively evaluated by overlaying the derived bone masks onto the corresponding MR images. The four common segmented regions between MR and CT were quantitatively assessed using the Dice coefficient and Hausdorff

Table 2. Quantitative comparison of downstream segmentation metrics (Dice coefficient and Hausdorff distance) between 3D registration networks (Voxelmorph, LapIRN, TransMorph) and an I2I translation network (MUNIT).

Metric: Dice coefficient (↑)						
	CT(Ref)	Voxelmorph	LapIRN	TransMorph	MUNIT	Proposed
Hips	0.705	0.731	0.690	0.713	0.716	**0.742**
Femurs	0.771	0.808	0.772	0.804	0.788	**0.816**
Spinal cord	0.402	**0.461**	0.458	0.457	0.163	0.402
Urinary bladder	0.554	0.566	0.511	0.646	0.437	**0.678**
Metric: Hausdorff distance (↓)						
	CT(Ref)	Voxelmorph	LapIRN	TransMorph	MUNIT	Proposed
Hips	33.46	11.67	27.90	11.51	25.43	**10.97**
Femurs	14.82	8.93	12.58	11.78	17.35	**7.82**
Spinal cord	12.19	10.75	**10.06**	10.25	15.74	10.45
Urinary bladder	27.50	22.66	22.46	19.32	36.28	**16.17**

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

distance. We noticed that there were notable errors in MR segmentation but reported the values for fair comparisons.

Results: Figure 7 displays representative images illustrating severe misalignment in the bone region between MR and CT. While registration methods (Fig. 7(a)) demonstrate some improvement, residual misalignment is still noticeable. MUNIT (the top-performing method among I2I approaches) also struggled to produce accurate segmentation masks (Fig. 7(b)). In contrast, the proposed pseudo-CT method delivers notably accurate bone segmentation masks, highlighting the potential benefits of our method. In the quantitative evaluation, the proposed method demonstrated superior performance in most anatomical structures, with a significant advantage in the urinary bladder (Table 2). However, the overall Dice coefficient was lower than expected due to errors in the total-segmentor for MR segmentation mask inference (see Sec. 2 of the Supp.). However, this shows supporting evidence for the qualitative results.

5.3. Ablation study

Evaluation of the DACA block: We evaluate the effectiveness of the core component of our method, the DACA block, by adjusting two key elements (see Fig. 4): the local patch size (P) in cross-attention, and the use of deformation fields, denoted as Deformation Field (DF). Without DF, vanilla cross-attention is applied. The evaluation metrics, GC, NMI, and FID, are consistent with those used in the Results section. As shown in Table 3, both P and DF are important elements in the proposed architecture.

Evaluation of loss functions: In the proposed method, two different types of loss functions are used for training. One is the loss between the input MR images, and the other is the loss between the reference CT images (see Sec. 3.2). We compare conventional loss functions with our proposed ones. Firstly, we compare the proposed contrastive loss for MR images ($\mathcal{L}_{\text{patchNCE}}$) with $\mathcal{L}_{\text{mind}}$, which utilizes the modality-independent neighborhood descriptor (MIND) loss [6], commonly used to preserve structural correspondence across modalities, as seen in sc-CycleGAN [25]. Additionally, we compare the conventional mean absolute error ($\mathcal{L}_1$) with the contextual loss ($\mathcal{L}_{\text{cx}}$) for the CT images. When using $\mathcal{L}_{\text{patchNCE}}$ instead of $\mathcal{L}_{\text{mind}}$ for the MR images, the performance in terms of GC and NMI was significantly better (Table 4). As shown, the combination of $\mathcal{L}_{\text{patchNCE}}$ and $\mathcal{L}_{\text{cx}}$ achieved the best results, with lower FID and higher GC and NMI.

6. Conclusion and discussion

Our study introduces a novel approach for generating aligned CT data with respect to MR. Results demonstrate the efficacy of our method compared to registration methods (in terms of alignment) as well as I2I methods and the registration + I2I method (in terms of textual preservation). We also show the potential of using our method for example to effectively perform bone segmentation in MR with generated aligned CT. We consider another future application to improve downstream MR-to-CT synthesis tasks by using the generated pseudo-CT as a pseudo label. This might be applicable in cases where precise alignment was difficult.

In this study, due to the absence of ground truth caused by the misalignment between paired MR and CT images, we utilized indirect metrics such as GC and NMI to assess the cross-modality alignment. In future work, we plan to directly validate the proposed method's performance using well-aligned datasets, such as full-body phantoms for MR and CT.

In the current scope of this study, our method is designed to operate on 2D slices rather than in 3D due to the GPU memory limitations from memory-intensive cross-attention computation. We plan to extend the framework to 3D in future studies, utilizing memory-efficient transformers such as flash-attention or other modifications. However, despite this limitation, our method, computed in 2D slices and then stacked, demonstrates superior alignment compared to 3D registration methods (Tables 1 and 2). This confirms the efficacy of our method, and we anticipate further improvements when 3D operation becomes feasible.

Table 3. Ablation study for the DACA block. **P** denotes the local patch size in cross attention. **DF** indicates whether the deformation field is used. Gray denotes the setup of the proposed method.

P	DF	GC (↑)	NMI (↑)	FID (↓)
1	✓	0.280	0.353	**2.1**
20	✓	0.399	0.385	18.6
28		0.410	0.365	54.5
28	✓	**0.434**	**0.406**	19.2

Table 4. Ablation study for the evaluation of loss functions

with MR		with CT				
$\mathcal{L}_{\text{mind}}$	$\mathcal{L}_{\text{patchNCE}}$	$\mathcal{L}_1$	$\mathcal{L}_{\text{cx}}$	GC(↑)	NMI(↑)	FID(↓)
✓			✓	0.362	0.301	**18.3**
	✓	✓		0.412	0.378	33.1
	✓		✓	**0.434**	**0.406**	19.2

Acknowledgements. This work was supported by the Korea Institute of Science and Technology (KIST) Institutional Program under Grant No. 2E33204 and 2E32981. This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (No. RS-2023-00243034).

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

References

[1] Daniel Rueckert et al. “Nonrigid registration using free-form deformations: application to breast MR images”. In: *IEEE transactions on medical imaging* 18.8 (1999), pp. 712–721.

[2] Dinggang Shen and Christos Davatzikos. “HAMMER: hierarchical attribute matching mechanism for elastic registration”. In: *IEEE transactions on medical imaging* 21.11 (2002), pp. 1421–1439.

[3] CC Parker et al. “Magnetic resonance imaging in the radiation treatment planning of localized prostate cancer using intra-prostatic fiducial markers for computed tomography co-registration”. In: *Radiotherapy and oncology* 66.2 (2003), pp. 217–224.

[4] Brian B Avants et al. “Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain”. In: *Medical image analysis* 12.1 (2008), pp. 26–41.

[5] Marc Modat et al. “Fast free-form deformation using graphics processing units”. In: *Computer methods and programs in biomedicine* 98.3 (2010), pp. 278–284.

[6] Mattias P Heinrich et al. “MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration”. In: *Medical image analysis* 16.7 (2012), pp. 1423–1435.

[7] Delia Ciardo et al. “Multimodal image registration for the identification of dominant intraprostatic lesion in high-precision radiotherapy treatments”. In: *The British Journal of Radiology* 90.1079 (2017), p. 20170021.

[8] Xun Huang and Serge Belongie. “Arbitrary style transfer in real-time with adaptive instance normalization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1501–1510.

[9] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. “Unsupervised image-to-image translation networks”. In: *Advances in neural information processing systems* 30 (2017).

[10] A Vaswani. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* (2017).

[11] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[12] Xiaohuan Cao et al. “Region-adaptive deformable registration of CT/MRI pelvic images via learning-based image synthesis”. In: *IEEE Transactions on Image Processing* 27.7 (2018), pp. 3500–3512.

[13] Yuta Hiasa et al. “Cross-modality image synthesis from unpaired data using cyclegan: Effects of gradient consistency loss and training data size”. In: *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. Springer. 2018, pp. 31–41.

[14] Xun Huang et al. “Multimodal unsupervised image-to-image translation”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 172–189.

[15] Hsin-Ying Lee et al. “Diverse image-to-image translation via disentangled representations”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 35–51.

[16] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. “The contextual loss for image transformation with non-aligned data”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 768–783.

[17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding”. In: *arXiv preprint arXiv:1807.03748* (2018).

[18] Guha Balakrishnan et al. “Voxelmorph: a learning framework for deformable medical image registration”. In: *IEEE transactions on medical imaging* 38.8 (2019), pp. 1788–1800.

[19] Adrian V Dalca et al. “Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces”. In: *Medical image analysis* 57 (2019), pp. 226–236.

[20] Salman UH Dar et al. “Image synthesis in multi-contrast MRI with conditional generative adversarial networks”. In: *IEEE transactions on medical imaging* 38.10 (2019), pp. 2375–2388.

[21] Cheng-Bin Jin et al. “Deep CT to MR synthesis using paired and unpaired data”. In: *Sensors* 19.10 (2019), p. 2361.

[22] Tony CW Mok and Albert CS Chung. “Large deformation diffeomorphic image registration with laplacian pyramid networks”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer. 2020, pp. 211–221.

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.

[23] Taesung Park et al. “Contrastive learning for unpaired image-to-image translation”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer. 2020, pp. 319–345.

[24] Fuzhi Yang et al. “Learning texture transformer network for image super-resolution”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5791–5800.

[25] Heran Yang et al. “Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN”. In: *IEEE transactions on medical imaging* 39.12 (2020), pp. 4249–4261.

[26] Qianye Yang et al. “MRI cross-modality image-to-image translation”. In: *Scientific reports* 10.1 (2020), p. 3753.

[27] Sayantan Bhadra et al. “On hallucinations in tomographic image reconstruction”. In: *IEEE transactions on medical imaging* 40.11 (2021), pp. 3249–3260.

[28] Lingke Kong et al. “Breaking the dilemma of medical image-to-image translation”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1964–1978.

[29] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[30] Shadab Momin et al. “CT-MRI pelvic deformable registration via deep learning”. In: *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 11598. SPIE. 2021, pp. 329–334.

[31] Jiezhang Cao et al. “Reference-based image super-resolution with deformable attention transformer”. In: *European conference on computer vision*. Springer. 2022, pp. 325–342.

[32] Junyu Chen et al. “Transmorph: Transformer for unsupervised medical image registration”. In: *Medical image analysis* 82 (2022), p. 102615.

[33] Runze Han et al. “Deformable MR-CT image registration using an unsupervised, dual-channel network for neurosurgical guidance”. In: *Medical image analysis* 75 (2022), p. 102292.

[34] Alessa Hering et al. “Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning”. In: *IEEE Transactions on Medical Imaging* 42.3 (2022), pp. 697–712.

[35] Miao Kang et al. “Dual-stream pyramid registration network”. In: *Medical image analysis* 78 (2022), p. 102379.

[36] Jiahao Lu et al. “Is image-to-image translation the panacea for multimodal image registration? A comparative study”. In: *Plos one* 17.11 (2022), e0276196.

[37] Jiacheng Shi et al. “Xmorpher: Full transformer for deformable medical image registration via cross attention”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 217–226.

[38] Ruicheng Feng et al. “Generating aligned pseudo-supervision from non-aligned data for image restoration in under-display camera”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 5013–5022.

[39] Joel Honkamaa et al. “Deformation equivariant cross-modality image synthesis with paired non-aligned training data”. In: *Medical Image Analysis* 90 (2023), p. 102940.

[40] Lin Tian et al. “GradICON: Approximate diffeomorphisms via gradient inverse consistency”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 18084–18094.

[41] Jakob Wasserthal et al. “TotalSegmentator: robust segmentation of 104 anatomic structures in CT images”. In: *Radiology: Artificial Intelligence* 5.5 (2023).

[42] Tugba Akinci D’Antonoli et al. “TotalSegmentator MRI: Sequence-Independent Segmentation of 59 Anatomical Structures in MR images”. In: *arXiv e-prints* (2024), arXiv–2405.

Authorized licensed use limited to: Yonsei Univ. Downloaded on February 28,2026 at 09:27:40 UTC from IEEE Xplore. Restrictions apply.