Cross-Modality Image Registration via Generating Aligned Image using Reference-Augmented Framework

Daniel Kim, Abdullah Shazly, Mohammed A. Al-masni, Dong-Hyun Kim, Kanghyun Ryu

***Abstract*— Aligning a pair of cross-modality images (e.g., MR-CT, CBCT-CT) is important, yet conventional approaches, including registration or Image-to-Image (I2I) translation methods often have limitations. To overcome these challenges, we introduce a "Register by Generation (RbG)" framework, a novel 2D deep learning approach designed to generate images that are structurally well-aligned with the fixed image while preserving the detailed intensity and contrast of the moving image, which we refer to as the reference image. Our approach operates in two sequential key stages: first, we employ a novel semi-global reference-augmented image synthesis network incorporating Patch Adaptive Instance Normalization (PAdaIN). This method leverages a down-sampled reference image to guide local adaptive synthesis, generating a more accurately aligned image with a reduced risk of hallucinations. In the second stage, we introduce a detailed refining reference-augmented network featuring a Deformation-Aware Cross-Attention (DACA) block, which aims to recover finer details and textures that may be missing from the initial stage. This unique component (DACA block) enables the transfer of corresponding relevant features from the reference image, effectively performing a "copy-and-paste" operation within the latent feature space. Additionally, we propose a novel combination of loss functions that enables self-supervised training on misaligned datasets, eliminating the need for pre-aligned data. We rigorously evaluate our method on multiple misaligned datasets using metrics focused on structural alignment and distributional consistency, demonstrating comprehensively superior performance. Furthermore, we test its robustness by simulating intentional misalignments in a well-aligned dataset. Additionally, experiments from a case study and downstream segmentation tasks highlight the broad applicability of our approach. Source code: https://github.com/danny4159/RbG_framework**

***Index Terms*— Cross-modality image registration, Reference-augmented generation, Self-supervised learning**

This work was supported in part by the Korea Institute of Science and Technology (KIST) Institutional Program under Grant 2E33204, in part by the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) under Grant RS-2023-00243034. (Corresponding authors: Mohammed A. Al-masni; Dong-Hyun Kim; Kanghyun Ryu.)

Daniel Kim, Dong-Hyun Kim are with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea (e-mail: danny4159@yonsei.ac.kr; donghyunkim@yonsei.ac.kr).

Abdullah Shazly, Mohammend A. Al-masni are with the Department of Artificial Intelligence and Data Science, College of AI Convergence, Sejong University, Seoul 05006, South Korea (e-mail: abdullahshazli@sju.ac.kr; m.almasani@sejong.ac.kr).

Kanghyun Ryu is with the Intelligence and Interaction Research Center, Korea Institute of Science and Technology, Seoul 02792, South Korea (e-mail: khryu@kist.re.kr).

I. Introduction

Cross-modality image registration, which achieves structural alignment between images of different modalities, is important for various applications. In neuroimaging, registration between multiple MRI sequences (e.g. T1, T2, FLAIR, and post-contrast T1) improves segmentation accuracy [1]. In addition, registration between MRI and CT images offers complementary information, with CT capturing bone and hard tissue structures and MRI for soft tissues [2]. For example, in radiation therapy for prostate cancer, CT can be used for dose planning and MRI for precise organ localization. Similarly, PET-CT registration can offer alignment of metabolic information from PET and anatomical information from CT [3], [4]. As such, the range of advantageous applications across various modalities continues to expand. Moreover, with the rising popularity of recent multi-modal deep learning models, such as Contrastive Language-Image Pre-training (CLIP) [5] and cross-attention transformers [6], achieving precise alignment across modalities will become more essential.

However, accurate alignment remains challenging, especially in deformable regions such as the abdomen, pelvis, and spine. Variability in patient posture [7]—such as CT or MRI acquired on a scanner versus CBCT with the patient sitting—introduces anatomical distortions in areas such as the spine, neck, and abdomen. Even with the patient scanned in the same position, the differences between MRI and CT scanners create further inconsistencies, as exact positioning cannot be replicated between machines [8], [9]. Physiological changes between scans, such as fluctuating organ fillings in the pelvis, varying lung air content, or changes in breathing patterns, introduce additional positional changes. These factors lead to complex deformations that are difficult to model, making registration inherently ill-posed and prone to blurred or erroneous output [10]–[12]. Finally, differences in image features and contrast between modalities complicate the correspondence; for example, MRI provides detailed soft tissue visualization, while CT may not offer the same clarity.

Registration methods that are developed to align two images have been vastly explored. A typical registration process starts with a fixed and moving image pair as input. A transformation

model estimates a deformation field, which indicates how the moving image should be warped to align with the fixed image. The consequent warped moving image is then compared with the fixed image using a similarity measure to quantify alignment accuracy. This similarity metric also serves as a loss function to optimize the transformation model through iterative updates.

To improve registration accuracy, prior research has primarily followed two complementary directions. The first focuses on enhancing transformation models with smoothness or diffeomorphic constraints to ensure topological consistency [13]. The second involves advancing transformation architectures through neural network-based approaches, such as convolutional encoder-decoders, vision transformers [14], and Mamba [15], [16], enabling faster and more flexible deformation prediction.

However, these methods are still unable to address the complexities of cross-modal medical image registration. Firstly, registration loss is typically computed globally, leading the transformation model to prioritize overall performance while overlooking local deformation errors. Usually, registration easily succeeds to align on a global scale, but often fails to accurately align small regions. Secondly, the constraints of the deformation model, such as diffeomorphic transformations, often break down in the presence of occlusions caused by physiological changes and variations in organ filling and air content [17]. Finally, the image metrics used to compute the loss function (e.g., NCC [18], MI [19], MIND [20]) used for cross-modal images are not fully effective; they primarily enforce correlations, common edges, or low-level features between the two images, which are suboptimal in intricate anatomical regions [21], [22].

In contrast, image-to-image (I2I) translation employs a network that transforms features from a source image—used as the sole input—to generate an image in the target domain. This approach is commonly applied to image synthesis tasks, such as MR-to-CT or CBCT-to-CT translation [23]–[25]. To improve I2I translation, prior studies [26]–[28] have introduced structural constraints to preserve anatomical consistency and adopted advanced architectures such as transformers and diffusion models [29]. Nevertheless, due to the inherently ill-posed nature of the source-to-target mapping and class imbalance in medical datasets, generation errors such as hallucination artifacts remain a persistent challenge [30], [31].

In this study, we take a fundamentally different approach from conventional registration methods by using ideas from Image-to-Image (I2I) translation techniques. Instead of aligning the moving image by warping with a deformation field, we propose to "generate" an image that is already structurally aligned with the fixed image (hereafter referred to as input image). However, unlike the conventional I2I framework, which relies solely on the input image, our method also utilizes the moving image (hereafter referred to as reference image) as an additional input to guide the generation process. This additional guidance improves the stability of the I2I framework and reduces the risk of hallucination artifacts.

We propose our novel framework and term it as "Register by Generation" (RbG) framework, implemented in a 2D to ensure computational efficiency. RbG takes both an input and a reference image as dual inputs, using the reference image to guide the generation process. It produces an output image that aligns structurally with the input image while accurately preserving the visual features (e.g., intensity and contrast) of the reference image.

We present our framework, emphasizing our unique approach of how the "misaligned" reference image can guide the network through two stages: an initial synthesis stage and a refinement stage. In the first stage, we employ a novel synthesis network that utilizes patch-based localized feature transfer (PAdaIN), guided by the reference image, to generate an initial output. This stage is conceptually inspired by style transfer [32]–[34], where the structural content is derived from the input image and the style from the reference image. The second stage refines this output using a Deformation-Aware Cross-Attention (DACA) block, which effectively aligns features between the input and reference images while addressing misalignment through a deformation field. Both stages leverage the reference image to achieve better alignment and details, drawing inspiration from prior approaches that combine I2I translation with registration [35], demonstrating that matching the style and contrast of the reference image facilitates the registration process.

We also present a novel self-supervised loss scheme that ensures structural alignment and distributional consistency, enabling the use of misaligned reference images as training data. This approach enhances the robustness of our framework and is useful for scenarios when dataset acquisition is difficult.

To validate our method's performance, we conducted extensive evaluations using three public datasets. The first dataset includes MR and CT pairs of pelvic images, which exhibit misalignment throughout the entire image. The second dataset consists of CBCT and CT image pairs in brain imaging, focusing on the cervical spine region where misalignment is often due to posture differences. The third dataset uses simulated T1 and T2 pairs to allow detailed retrospective validation of our approach. Additionally, we performed a comprehensive comparison with various conventional methods, including registration and I2I translation. An ablation study further proved the effectiveness of our novel components, such as PAdaIN and the DACA block, demonstrating their superiority over standard AdaIN-based global guidance and conventional cross-attention blocks. Finally, we demonstrate the utility of our method by transferring CT features to MR coordinates, effectively preserving detailed bone structures and improving segmentation accuracy compared to conventional registration methods. This performance was also quantitatively validated by using masks generated on MR images with a third-party open-source segmentation tool, TotalSegmentator.

II. RELATED WORKS

A. Registration Methods

Optimization-based methods estimate the deformation field through iterative optimization and can be broadly divided into intensity-based and feature-based approaches. In intensity-based methods, the intensity differences between images are

directly compared. These methods maximize similarity measures such as mutual information (MI), modality-independent neighborhood descriptor (MIND), and generalized cross-correlation (GCC) [19], [20]. However, these similarity measures had limitations in capturing the spatial information of anatomical structures [36]. In feature-based methods, techniques such as Scale-Invariant Feature Transform (SIFT) are used to detect invariant features and identify corresponding points between images. While effective in some cases, their performance was limited by the complexity and variability of medical images [37], [38].

Deep learning (DL)-based methods offer significant advantages over optimization-based methods by addressing high computational demands and reducing the risk of falling into local optima [36]. These methods are broadly categorized into mono-modality and cross-modality registration, where mono-modality techniques are designed for images with the same contrast, while cross-modality approaches handle images with differing contrasts.

Mono-modality registration estimates the deformation field between images of the same modality. For instance, VoxelMorph [13] utilized a CNN-based network with dual inputs to perform spatial correspondence, often using mean squared error (MSE) loss to measure similarity between the warped (transformed) moving image and the fixed reference image, effectively minimizing intensity differences for accurate alignment. These approaches enable straightforward unsupervised training, as training pairs can be easily generated by shifting images to create synthetic moving and fixed image pairs. TransMorph [39] incorporated a self-attention mechanism, improving its ability to handle long-range feature correspondences. Additionally, methods like GradICON [40] enhanced regularization of displacement fields, achieving approximately diffeomorphic mappings with inverse consistency loss rather than traditional smooth loss functions [41].

Cross-modality registration requires a more complex understanding of cross-modal features. This approach often relies on similarity metrics like MI [19] or MIND [20], which are less optimal than MSE loss for mono-modality registration because they do not directly measure pixel-wise intensity differences. For example, Patch-RegNet [42] performed registration between MR and CT images using a transformer-based, multi-resolution approach, while Attention-Reg [43] applied contrastive learning to calculate cross-feature correlations between MR and Transrectal Ultrasound (TRUS) images. Despite these advances, identifying shared features across different modalities remains challenging and can lead to inaccuracies in deformation estimation [7], with high dependency on data quality and availability [42], [43].

B. Image-to-Image (I2I) Translation

Unsupervised I2I translation aimed to transform images between different domains without requiring paired datasets. This approach can serve as an alternative to registration methods by generating a consistent contrast image rather than working with misaligned images.

Following the success of CycleGAN [44], numerous extension methods based on this framework were developed. To address CycleGAN's limitation of structural inconsistency between input and generated images, methods like ShapeGAN [26], GCGAN [27], and SCGAN [28] introduced direct constraints with the input images to better preserve structural integrity. CycleSGAN [45] further improved anatomical preservation by incorporating an additional discriminator that utilized annotation labels, functioning as a three-player GAN model. Additionally, SynDiff [29], recent work in this field, achieved conditional translation through a multistep reverse diffusion process, significantly improving stability and fidelity.

Disentangled representation-based approaches, which separated images into content and style codes to generate new images, were also widely applied. Notable works include UNIT [46], MUNIT [47], and DRIT [48]. Extensions of these methods in the field of medical imaging, integrated constraints to ensure consistency in anatomical structure [49]–[51]. While similar to I2I translation, this approach aligned more closely with our method by using the style of one contrast (moving) and the content of another (fixed), enabling a more effective blending of cross-modal features.

However, these methods were susceptible to generating errors and hallucinations, particularly due to the under-representation of certain classes, such as contrast agents and tumors [30], [31]. They often struggled to preserve the unique visual features of individual patient reference images. Ensuring the preservation of these critical features in reference images was a primary objective of our research.

C. Integrated Approach: Registration + I2I Translation

The integrated approach sequentially combined registration and I2I translation to enhance cross-modality registration performance. For instance, RegGAN [35], Hémon et al. [52], and OTMorph [7] first performed I2I translation to align the image domains and then conducted registration. OTMorph [7], in particular, employed deterministic Neural Optimal Transport [53] for I2I translation, focusing on preserving anatomical structures. Conversely, GC-Reg [36] estimated the deformation field in a cross-modality context before performing I2I translation, employing a reconstruction loss during training. However, these sequential methods were often susceptible to errors introduced during the I2I translation step, such as hallucinations or the loss of high-frequency details [42]. In contrast, our approach integrates registration directly within the I2I architecture, using cross-attention modules to guide the feature transfer, thus ensuring more effective and cohesive feature alignment between images.

D. Reference-Guided Image Reconstruction (Ref-Recon)

This study is inspired by the recent introduction of Ref-Recon, which aims to enhance image reconstruction by using high-quality, semantically related reference images as additional guidance. In natural image super-resolution, methods such as TTSR [54] and DATSR [55] have employed Transformer mechanisms to integrate reference features. In medical imaging, Ref-Recon techniques have also recently been applied to tasks such as undersampling [56]–[59] and

resolution enhancement [60], [61], using multi-contrast images or other volumes within the same contrast as reference images to improve reconstruction quality. Although our study is inspired by these recent advancements, we uniquely apply these concepts to the registration problem. Rather than focusing on image quality enhancement alone, our approach emphasizes the accurate transfer of information across misaligned cross-modalities, addressing the challenges of modality translation in a novel way.

III. METHOD

A. Problem Statement

The objective of the proposed "Register by Generation" (RbG) framework is to address the limitations of conventional image registration and Image-to-Image (I2I) translation methods. Specifically, RbG generates images that structurally align with the input (x) image, while preserving the intensity and contrast features of a potentially misaligned reference (y) image. The framework employs a two-stage process to ensure both structural alignment and feature retention, especially in scenarios with complex anatomical deformations and varying image modalities.

$$\min_{\hat{y}} \left[\mathcal{L}_{\text{struct}}(x, \hat{y}) + \mathcal{L}_{\text{feat}}(y, \hat{y})\right] \tag{1}$$

Mathematically, we aim to generate an output ($\hat{y}$) image that minimizes structural and feature discrepancy metrics with respect to both the input (x) and the reference (y). This optimization is guided by a combination of structural alignment ($\mathcal{L}_{\text{struct}}$) and contextual feature ($\mathcal{L}_{\text{feat}}$) losses. The overarching goal is to produce images that satisfy the Equation (1).

B. Overview of Our Method

The key challenge in our RbG framework is the accurate transfer of features from a "misaligned" reference (y) image. To address this, we adopt a two-stage framework. Stage 1 synthesizes an initial output ($\acute{y}$) in the same modality as the reference (y) image, thereby reducing the modality gap. This enables Stage 2 to perform fine-grained refinement guided by a uni-modality registration, which is more stable than cross-modality alignment (see Fig. 1(a)). For computational efficiency, all components are implemented in 2D.

1) Stage 1 (Initial Synthesis): We generate an initial output $\acute{y} = G(x, y)$ using a Synthesis Network (G) that employs Patch Adaptive Instance Normalization (PAdaIN) to transfer localized features from the reference (y) image to align with the input (x) image.

2) Stage 2 (Feature Refinement): The initial output ($\acute{y}$) is refined using a Deformation-Aware Cross-Attention (DACA) mechanism. A deformation field $\phi = R(\acute{y}, y)$, estimated by a pre-trained registration network (R), guides the features fusion from y into $\acute{y}$ to produce the final aligned output ($\hat{y}$).

C. Details of Stage 1

The objective of this stage is to generate an initial output ($\acute{y}$) aligned with the input (x) image using the synthesis network (G). To improve synthesis quality, we incorporate guidance from the reference (y) image through feature transfer. However, due to the inherent misalignment between the input-reference image pair, a naive feature transfer may result in incorrect correspondence. Inspired by the spatial adaptive style transfer method [33], we compress and transfer the spatial features of the reference (y) image in a localized manner. To this end, the proposed Patch Adaptive Instance Normalization (PAdaIN) is applied across multiple scales, enabling spatially-aware feature transfer.

The patch-wise contrast and style of the reference (y) image can be reflected during the PAdaIN process. The concept is similar to the AdaIN [32] mechanism. However, unlike conventional AdaIN-based studies that perform global feature transfer, our approach focuses on transferring features at a patch level, effectively capturing intricate details.

In PAdaIN, the reference (y) image is spatially downsampled by a factor of n to aggregate information at a patch level. This coarse representation serves as a bottleneck that discards unreliable high-frequency details while retaining sub-local structures crucial for anatomical fidelity. Then upsampling is applied to match the spatial size of the target feature map ($F_x^{(l)}$) extracted from the input (x) image at the l-th layer of the synthesis network (G) (in Eq. (2)), resulting in $y^{\downarrow(l)}$. Bilinear interpolation is used to preserve the continuity of feature representation.

$$y^{\downarrow(l)} = \text{Up}\left(\text{Down}\left(y, \text{factor} = n\right), \text{spatial size}\left(F_x^{(l)}\right)\right) \tag{2}$$

Additionally, two Multi-Layer Perceptrons (MLPs) adjust $y^{\downarrow(l)}$ to match the channel dimensions of $F_x^{(l)}$, generating the gamma ($\gamma^{(l)}$) and beta ($\beta^{(l)}$) (in Eq. (3)).

$$\gamma^{(l)} = \text{MLP}_\gamma(y^{\downarrow(l)}), \quad \beta^{(l)} = \text{MLP}_\beta(y^{\downarrow(l)}) \tag{3}$$

After applying instance normalization to $F_x^{(l)}$, where the mean $\mu(F_x^{(l)})$ and standard deviation $\sigma(F_x^{(l)})$ are computed across the spatial dimensions for each channel, scaling and shifting are performed using the gamma and beta, enabling the statistical feature transfer (in Eq. (4)).

$$\text{PAdaIN}(F_x^{(l)}, y) = \gamma^{(l)} \cdot \frac{F_x^{(l)} - \mu(F_x^{(l)})}{\sigma(F_x^{(l)})} + \beta^{(l)} \tag{4}$$

As demonstrated in StyleGAN [62], Gaussian noise is added to each layer of the synthesis network (G) to enhance the representation of fine features. The proposed PAdaIN is applied to multi-scale features and performs local transfer, effectively handling not only the misalignment of the reference (y) image but also complex feature representations.

D. Details of Stage 2

The purpose of this stage is to recover finer details that may be missing from the initial output ($\acute{y}$) from Stage 1, by refining features from the misaligned reference (y) image. This is achieved using the proposed Deformation-Aware Cross-Attention (DACA) block, where each pixel of the initial output ($\acute{y}$) is fused with the neighborhood of the corresponding pixel from the reference (y) image through cross-attention. Here, the

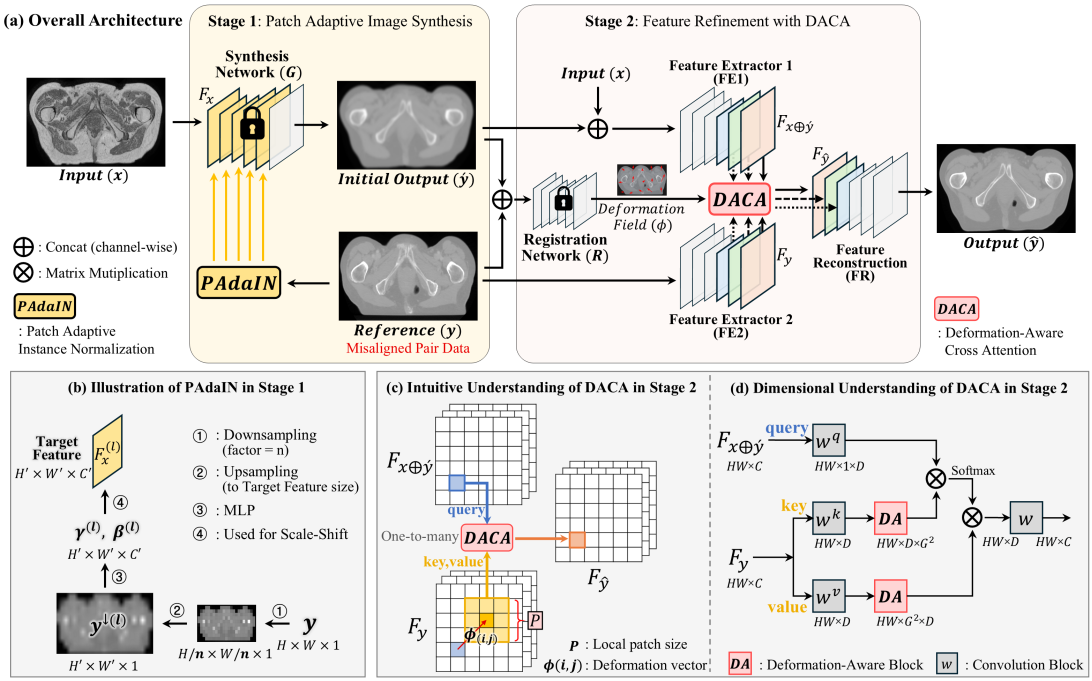

Fig. 1. Visualization of the proposed framework and its key components. (a) The overall architecture of the framework: Stage 1 generates the "Initial Output" while Stage 2 refines it using the DACA block, which fuses features from the "Reference" image to produce the final "Output." (b) The PAdaIN mechanism used in Stage 1, enabling localized feature transfer. (c, d) Detailed illustration of the DACA block in Stage 2.

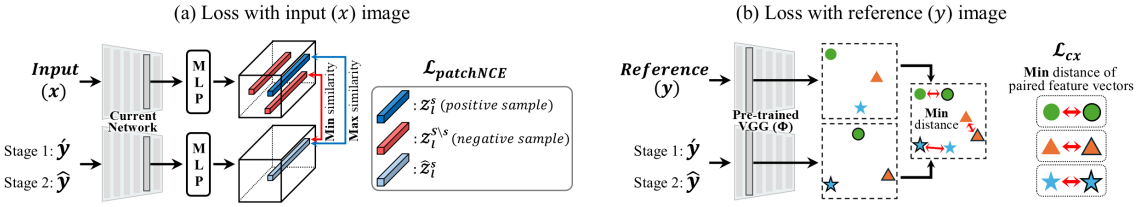

Fig. 2. Visualization of the loss functions used for training in both Stage 1 and Stage 2. The combination of the two losses, shown in (a) and (b), aims to generate an output that aligns with the input image while preserving the features of the reference image.

"corresponding pixel" refers to the warped coordinate obtained via the deformation field (ϕ).

Firstly, the deformation field (ϕ) is estimated by a pre-trained registration network (R). The initial output ($\acute{y}$) and reference (y) image serve as an input pair for mono-modality registration. Before training Stage 2, the registration network (R) is pre-trained with the same input pair ($\acute{y}$ and y) and then frozen, as its objective is inherently different from the objective of Stage 2, which is to align the output ($\hat{y}$) with the input (x) image. This is the fundamental difference between standard registration approaches that iteratively optimize the deformation field estimation and our method. For fast estimation, we adopted a learning-based method and followed the widely used Voxelmorph [13] network architecture. Note, however, that our framework is flexible and does not depend on any specific network; any off-the-shelf, non-learning-based method—such as numerically optimized approaches—can be used to provide the required deformation field.

$$\phi(i,j) = R(\acute{y}, y) \tag{5}$$

Next, two U-Net-based convolutional feature extractors (FE1 and FE2) are employed. FE1 takes a dual-channel input of the input (x) image and the initial output ($\acute{y}$) to obtain $F_{x\oplus\acute{y}}$, while FE2 receives the single-channel reference (y) image to extract F_y . Incorporating the input (x) image into FE1 is beneficial, as it is structurally aligned with the initial output ($\acute{y}$) and less sensitive to modality differences due to contrastive training (see Section III-E-1). This provides complementary structural information to guide feature extraction. Then, following the standard transformer, independent convolution blocks are used for channel projection, obtaining "query" from $F_{x\oplus\acute{y}}$, and "key" and "value" from F_y.

One-to-many cross-attention is employed to fuse the query with the key and value (see Fig.1(c)). In this process, the warped coordinates are applied to the "key" and "value" to ensure accurate feature alignment (as in Eq. (6)). One-to-many matching occurs between a pixel of "query" and the surrounding local patch of the corresponding pixel of the "key" and "value", enabling the fusion of localized features from the reference (y) image. This one-to-many matching helps mitigate feature misalignment caused by minor errors in the registration network (R), particularly in anatomically intricate regions. This design is performed at multiple scales.

$$(i', j') = (i + \phi_w(i,j), j + \phi_h(i,j)) \tag{6}$$

$$F_{\hat{y}}(i,j) = \sum_i \sum_j \text{Attention}(F_{x\oplus\hat{y}}(i,j), F_y(i'+p, j'+p)) \quad (7)$$
$$\text{for all } p \in \{-\lfloor P/2 \rfloor, \ldots, \lfloor P/2 \rfloor\}, p \in \mathbb{Z}$$

The fused features ($F_{\hat{y}}$) serve as the encoder representation of the convolutional Feature Reconstruction (FR) network, which is decoded to generate the final output ($\hat{y}$). As demonstrated in the results section, the proposed DACA block successfully produces an output that aligns with the input (x) image while incorporating patient-specific features from the reference (y) image.

E. Proposed Loss Function for Self-Supervised Learning Scheme

The objective of our network is to achieve two goals simultaneously for the output image. The first is to ensure that it is aligned with the input (x) image, and the second is to incorporate the features of the misaligned paired reference (y) image. To accomplish these objectives, we propose a combination of loss functions that is used for training both Stage 1 and Stage 2.

1) Structural alignment loss with input image (Fig.2(a)): To ensure structural alignment with the input (x) image, we employ a Patch-based Noise Contrastive Estimation (PatchNCE) based contrastive loss. This loss function maximizes mutual information by preserving local structures between two images from different domains. The loss function is defined as follows:

$$\mathcal{L}^{(l)}_{\text{patchNCE}}(\tilde{y}, x) = \mathbb{E}\left[\sum_{s=1}^{S_l} \ell(\hat{z}_l^s, z_l^s, z_l^{S\setminus s})\right], \quad (8)$$

$$\text{where } \tilde{y} = \begin{cases} \acute{y}, & \text{for Stage 1} \\ \hat{y}, & \text{for Stage 2} \end{cases}$$

We define the "feature vectors" as multiple local patches sampled from the feature map of the current network—the synthesis network G in Stage 1, and FE1 and FE2 in Stage 2—each vectorized via an MLP. Specifically, $\hat{z}_l^s$ denotes the feature vector derived from the output ($\tilde{y}$) image, while z_l^s and $z_l^{S\setminus s}$ are from the input (x) image. The term ℓ maximizes the cosine similarity between $\hat{z}_l^s$ and its spatially corresponding vector z_l^s, while minimizes the similarity to spatially non-matching vectors $z_l^{S\setminus s}$. This loss ensures alignment and penalizes different regions in the feature space, thereby improving structural alignment between cross-modality images.

Unlike the original PatchNCE loss, in Stage 2, our method employs a pre-trained VGG-19 network (Φ) for the feature extraction instead of the current network (FE1 and FE2), as the complexity of this stage prevented convergence with the current network. To focus on structural information, the high-level feature representations from the "conv4_2" and "conv5_4" layers of the VGG-19 network were utilized.

2) Feature preserving loss with reference image (Fig.2(b)): In real-world scenarios, obtaining an aligned reference (y) image is often infeasible. Therefore, we use the misaligned reference (y) image as a proxy label and adopt a self-supervised learning approach.

The use of pixel-wise losses such as MSE or MAE is not suitable for capturing features in misaligned images. Although perceptual loss [63] applied at the feature level can handle minor misalignments, it tends to lead to errors in cases of more significant misalignment. Therefore, we employ a contextual loss [64] that is robust to misalignment. This loss matches the most similar pixel pairs at the feature level, regardless of pixel-wise correspondence, and minimizes their cumulative distance. The loss function is defined as follows:

$$\mathcal{L}^{(l)}_{cx}(\tilde{y}, y) = \mathbb{E}\left[-\frac{1}{N}\sum_i \min_j \psi(\Phi^{(l)}(\tilde{y})_i, \Phi^{(l)}(y)_j)\right], \quad (9)$$

where $\tilde{y}$ represents the output image, which is $\acute{y}$ in Stage 1 and $\hat{y}$ in Stage 2. Here, $\Phi^{(l)}(\cdot)$ denotes the feature map extracted from layer l of the pre-trained VGG-19 network (Φ), and ψ measures the cosine similarity. The $\min$ operation pairs each feature vector i from the output ($\tilde{y}$) image with the most similar feature vector j from the reference (y) image. By minimizing the cumulative distance across all such feature vector pairs, this approach helps to mitigate misalignment between the images. In our experiments, multi-scale feature maps from the "conv1_2", "conv2_2" and "conv3_2" layers were utilized to effectively capture the contrast and texture of the reference (y) image, focusing on low-level features.

To enhance high-quality image generation, we used a GAN-based discriminator loss, encouraging the generation of images indistinguishable from the reference (y) image. We specifically employed the patch-based Least Squares GAN (LSGAN), which stabilizes training and minimizes vanishing gradient issues, resulting in more realistic image generation.

3) Total loss: The total loss is calculated as follows, with the weighting balance depending on the task.

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

IV. Experiments

A. Datasets and Preprocessing

All preprocessing procedures followed commonly adopted practices, with reference to the preprocessing strategies described in [65]–[67].

1) Misaligned dataset:

*a) **MR-CT pelvis dataset**:* This dataset from the SynthRAD2023[1] Grand Challenge includes MR and CT image pairs acquired from two different institutions. The CT images were resampled to $1 \times 1 \times 2.5\,\text{mm}^3$, and rigid registration was applied to align with the MR images. However, despite this alignment, significant misalignment between the MR and CT pairs remains. The MR images were captured using T1-weighted gradient echo sequences at 1.5T or 3T, and inter-scanner intensity variations were standardized using Nyúl histogram matching [67]. CT images underwent Hounsfield Unit (HU) value clipping to the 5th and 95th percentiles to remove intensity outliers and normalize the intensity range prior to training. Both MR and CT images were resized and padded to maintain a consistent resolution of 384×320 while preserving aspect ratios. Each patient's data was normalized

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

to a $[-1, 1]$ intensity range. The dataset contains 165 patients, split into 112 for training, 18 for validation, and 35 for testing.

*b) **CBCT-CT brain dataset**:* This dataset, also part of the SynthRAD2023 Grand Challenge, consists of CBCT and CT image pairs from three different institutions. Both CBCT and CT images were resampled to an isotropic resolution of $1 \times 1 \times 1\,\text{mm}^3$, and rigid registration was applied between them. Despite the registration, there remains residual misalignment between the paired images. HU value clipping was applied to both CBCT and CT images at the 1st and 99th percentiles. All images were then cropped and padded to a resolution of 196×196, followed by normalization to the $[-1, 1]$ range. The dataset includes 180 patients, with 126 used for training, 18 for validation, and 36 for testing.

2) Well-aligned dataset:

*a) **T1-T2 brain dataset**:* This dataset, sourced from the IXI[2] dataset, was used to enable retrospective validation on well-aligned data. It consists of T1 and T2 MR image pairs collected from 40 healthy subjects, with each pair comprising 91 slices. T1 images were acquired with TR/TE = 9.8/4.6 ms, flip angle = 8°, and volume size of $256 \times 256 \times 91$ in sagittal orientation, while T2 images were acquired with TR/TE = 8178/100 ms, flip angle = 90°, with the same volume size in axial orientation. To reduce inter-scanner intensity variation in MR images, Nyúl histogram matching was applied. The images were resampled to a resolution of $0.94 \times 0.94 \times 1.2\,\text{mm}^3$ and pre-aligned, and all intensity values were normalized to the range $[-1, 1]$. The intentional misalignment was applied to the training data to evaluate misalignment robustness, specifically using TorchIO[3] functions. These transformations included random in-plane rotations from -5° to +5°, translations from -10 mm to +10 mm, and 3D deformations ranging from -5 mm to +5 mm, which were also applied in a through-plane. The dataset is divided into 25 subjects for training, 5 for validation, and 10 for testing.

B. Implementation Details

1) Stage 1 (Synthesis network): The synthesis network uses normal initialization for all convolution layers and maintains 256 channels throughout. The Adam optimizer is applied with a learning rate and weight decay of 0.0001, and betas of 0.5 and 0.999. A multi-step scheduler is used to reduce the learning rate by 0.1 at the 20th and 30th epochs. The sampling function for the patchNCE loss is initialized using Xavier initialization. The PAdaIN method is applied with a downsampling factor of 32 (see Table V), which is dataset-dependent. The loss weights for $\lambda_{\text{patchNCE}}$, λ_{cx}, and λ_{gan} are set at 5, 5, and 1, respectively. A batch size of 8 was used to maximize memory utilization, and the model was trained for 100 epochs, requiring 60 hours on a single NVIDIA RTX A5000 GPU for the Pelvis MR-CT pair dataset.

2) Registration network: For the registration task, 2D registration is learned between the reference image and the initial output from the synthesis network in a uni-modality setting. The network is based on a convolutional architecture using Voxelmorph [13] as the baseline. The Adam optimizer is used with default settings, and a multi-step scheduler reduces the learning rate by 0.1 at the 20th and 30th epochs. The loss function combines MSE loss with a weight of 1 and a smooth loss, applied to regularize gradients in the aligned image, with a weight of 0.5. The network was trained with a batch size of 10 for 100 epochs, taking 10 hours on a single NVIDIA RTX A5000 GPU for the Pelvis MR-CT pair dataset.

3) Stage 2: In Stage 2, the pre-trained synthesis network from Stage 1, along with the registration network, is fixed. For cross-attention, we use two heads and a local patch size of 28 (see Table VI). Both the feature extractor (FE) and feature reconstruction (FR) employ a U-Net-like architecture composed of convolutional layers, initialized with normal initialization. The Adam optimizer is used with default settings, and the learning rate is reduced by 0.1 at the 20th and 30th epochs using a multi-step scheduler. The loss weights for $\lambda_{\text{patchNCE}}$, λ_{cx}, and λ_{gan} are set to 1, 10, and 1, respectively, as Stage 2 focuses more on capturing feature information from the reference image compared to Stage 1. To prevent the output image from aligning with the reference image when computing contextual loss, random translations within [-10, 10] are applied to the reference image. A batch size of 4 was used to maximize memory utilization, with training conducted for 100 epochs, taking 64 hours on a single NVIDIA RTX A5000 GPU for the Pelvis MR-CT pair dataset.

C. Comparison Methods and Evaluation Metrics

1) Comparison methods: As introduced in Section II, we compare three methods for multi-modality image alignment: Registration, I2I translation, and an integrated approach that combines I2I translation and registration.

In the registration approach, to reduce the complexity of training across different modality image pairs, we first perform I2I translation to generate synthesized images, followed by uni-modality registration between the synthesized image and the reference image. MUNIT, which consistently delivers superior results (see Table I), is used for the I2I translation network. We also employ previous state-of-the-art (SOTA) models such as SyN [68], Voxelmorph [13], LapIRN [69], GradICON [40], and TransMorph [39] for comparison.

For unsupervised I2I translation, we employ models such as CGAN [44], SCGAN [28], UNIT [46], and MUNIT [47].

For the integrated method, we use RegGAN [35] as a baseline, which dual-trains I2I translation and registration simultaneously. Additionally, inspired by [70], we apply regularization to ensure consistent results when reversing the order of I2I translation and registration, improving performance.

2) Evaluation metrics:

*a) **For misaligned dataset**:* For misaligned datasets such as MR-CT and CBCT-CT, where aligned ground truth is unavailable, we utilize the following evaluation metrics.

First, to assess structure alignment with the input (x) image, we use Gradient Correlation (GC) based on canny edge detection and Normalized Mutual Information (NMI). Second, to evaluate the distribution consistency (i.e., intensity/contrast similarity) with the reference (y) image, we

[2] https://brain-development.org/ixi-dataset/
[3] https://torchio.readthedocs.io/

TABLE I

EVALUATION OF THE PROPOSED METHOD AGAINST COMPARISON METHODS ON THE MR-CT PELVIS DATASET. GC AND NMI EVALUATE STRUCTURE ALIGNMENT WITH THE INPUT IMAGE, WHILE FID AND KID ASSESS DISTRIBUTIONAL CONSISTENCY. DARK GRAY INDICATES THE TOP-PERFORMING METHOD, WHILE LIGHT GRAY REPRESENTS THE SECOND-BEST. STATISTICAL SIGNIFICANCE AGAINST THE PROPOSED METHOD: $^{**}(p < 0.001)$, $^{*}(p < 0.05)$, $^{-}(p \geq 0.05)$.

Types	Methods	Input as ground truth GC ↑	Input as ground truth NMI ↑	Reference as ground truth FID ↓	Reference as ground truth KID ↓ ($\times 10^{-2}$)	No ground truth Sharpness ↑
Reference img	-	0.23 ± 0.05	0.31 ± 0.05	-	-	87.1 ± 41.5
Registration	SyN	0.29 ± 0.05 **	0.45 ± 0.05 **	9.9	0.56 ± 0.32 -	55.0 ± 18.4 **
	Voxelmorph (2D)	0.30 ± 0.06 **	0.37 ± 0.06 **	14.0	0.72 ± 0.31 **	74.3 ± 25.6 **
	Voxelmorph (3D)	0.28 ± 0.08 **	0.37 ± 0.05 **	32.9	1.71 ± 0.34 **	88.5 ± 21.5 **
	LapIRN (3D)	0.30 ± 0.08 **	0.37 ± 0.05 **	33.8	1.96 ± 0.37 **	71.9 ± 14.2 **
	GradICON (2D)	0.28 ± 0.06 **	0.36 ± 0.05 **	21.6	1.12 ± 0.35 **	61.7 ± 16.2 **
	GradICON (3D)	0.26 ± 0.06 **	0.36 ± 0.04 **	26.3	1.15 ± 0.35 **	85.0 ± 19.8 **
	TransMorph (3D)	0.31 ± 0.08 **	0.38 ± 0.05 **	35.7	2.44 ± 0.48 **	106.8 ± 20.8 **
I2I Translation	CGAN	0.15 ± 0.05 **	0.33 ± 0.03 **	60.6	4.45 ± 0.74 **	66.2 ± 19.5 **
	SCGAN	0.11 ± 0.04 **	0.29 ± 0.03 **	123.9	14.74 ± 1.46 **	79.5 ± 28.7 **
	UNIT	0.45 ± 0.06 **	0.36 ± 0.03 **	32.2	1.58 ± 0.34 **	60.9 ± 19.8 **
	MUNIT	0.34 ± 0.06 **	0.33 ± 0.03 **	28.9	1.34 ± 0.35 **	80.6 ± 24.3 **
Integrated	RegGAN	0.38 ± 0.05 **	0.04 ± 0.04 **	59.1	4.42 ± 4.21 **	89.1 ± 14.6 **
Proposed	RbG (Proposed)	0.43 ± 0.07	0.41 ± 0.04	16.2	0.50 ± 0.30	99.8 ± 19.6

TABLE II

EVALUATION OF THE PROPOSED METHOD AGAINST COMPARISON METHODS ON THE CBCT-CT BRAIN DATASET.

Types	Methods	Input as ground truth GC ↑	Input as ground truth NMI ↑	Reference as ground truth FID ↓	Reference as ground truth KID ↓ ($\times 10^{-2}$)	No ground truth Sharpness ↑
Reference img	-	0.50 ± 0.15	0.43 ± 0.06	-	-	121.4 ± 85.1
Registration	SyN	0.35 ± 0.13 **	0.35 ± 0.12 **	13.1	0.53 ± 0.15 **	71.5 ± 57.4 **
	Voxelmorph (2D)	0.50 ± 0.15 **	0.42 ± 0.05 **	7.6	0.30 ± 0.19 -	76.2 ± 53.3 **
	Voxelmorph (3D)	0.50 ± 0.15 *	0.42 ± 0.05 **	10.1	0.47 ± 0.21 **	75.9 ± 52.7 *
	LapIRN (3D)	0.49 ± 0.15 **	0.41 ± 0.05 **	23.7	1.22 ± 0.38 **	94.2 ± 74.0 *
	GradICON (2D)	0.42 ± 0.14 **	0.44 ± 0.14 -	16.4	0.53 ± 0.15 **	29.4 ± 20.7 **
	GradICON (3D)	0.43 ± 0.15 **	0.42 ± 0.15 *	11.2	0.52 ± 0.17 **	150.7 ± 168.2 **
	TransMorph (3D)	0.49 ± 0.15 **	0.41 ± 0.05 **	18.2	0.73 ± 0.20 **	101.0 ± 76.2 -
I2I Translation	CGAN	0.56 ± 0.16 -	0.46 ± 0.06 -	15.4	1.28 ± 0.27 **	120.3 ± 79.0 -
	SCGAN	0.56 ± 0.16 -	0.42 ± 0.05 **	17.0	1.57 ± 0.40 **	119.0 ± 61.8 -
	UNIT	0.53 ± 0.15 -	0.43 ± 0.05 **	24.1	1.15 ± 0.40 **	100.2 ± 46.6 -
	MUNIT	0.55 ± 0.15 -	0.44 ± 0.05 *	18.9	1.48 ± 0.41 **	77.9 ± 49.1 **
Integrated	RegGAN	0.19 ± 0.06 **	0.19 ± 0.03 **	416.1	N/A	N/A
Proposed	RbG (Proposed)	0.57 ± 0.16	0.46 ± 0.06	7.6	0.34 ± 0.21	115.4 ± 80.2

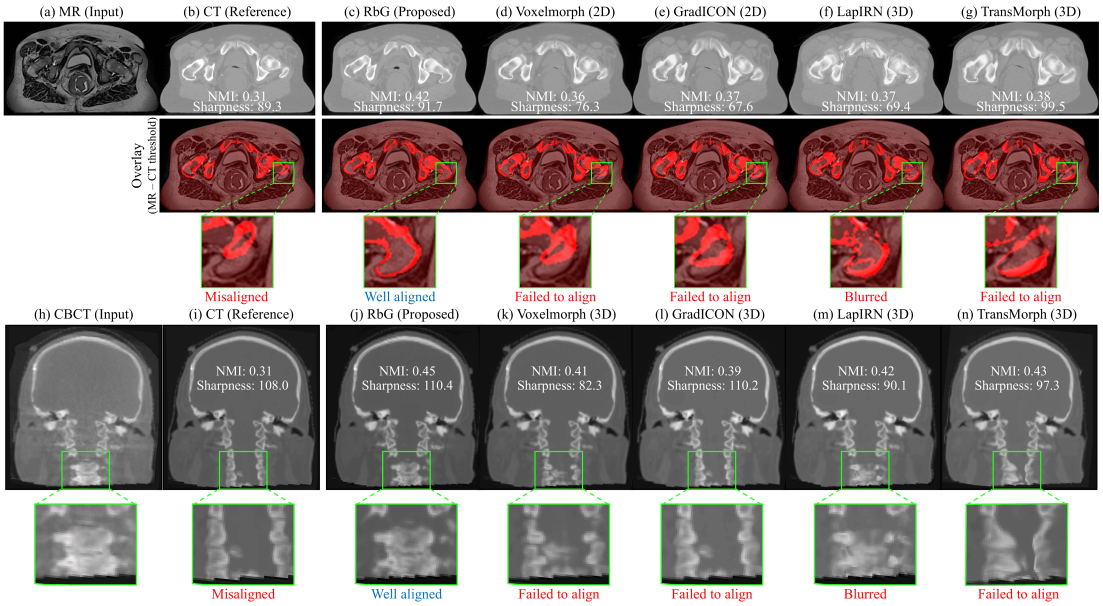

Fig. 3. Comparative analysis of registration methods on MR-CT (2D, except LapIRN) and CBCT-CT (3D). Higher NMI values indicate better structural alignment with the input image.

apply Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). Sharpness is evaluated using the variance of the Laplacian metric, highlighting its importance in medical imaging, as confirmed through downstream segmentation tasks in the results section. Wilcoxon signed-rank tests [71] were conducted for all metrics except FID, which lacks per-sample scores.

*b) **For well-aligned dataset**:* For well-aligned dataset, such as T1-T2, we use pixel-wise metrics for direct evaluation. These include the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and the high-level feature-based Learned Perceptual Image Patch Similarity (LPIPS) metric to assess image quality. Wilcoxon signed-rank tests were also applied to these metrics.

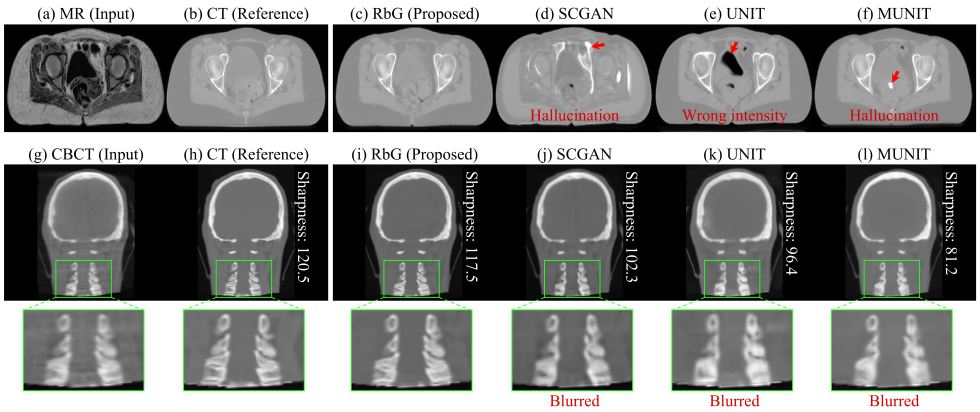

Fig. 4. Comparative Analysis of I2I Translation Methods on MR-CT Pelvis and CBCT-CT Brain Datasets.

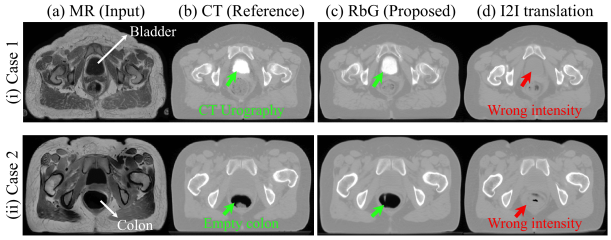

Fig. 5. Case study on the MR-CT Pelvis Dataset, comparing the proposed method with MUNIT. (i) evaluates feature preservation in contrast-enhanced areas, and (ii) examines underrepresented classes like empty colon regions.

V. RESULTS AND DISCUSSION

A. Experiment on the Misaligned Dataset

1) Quantitative and qualitative evaluation: In this study, we focus on datasets with severe input-reference image misalignment, specifically the MR-CT pelvis and CBCT-CT brain datasets. Due to the absence of aligned labels, we conduct indirect evaluations using various metrics such as GC, NMI, FID, and KID, aiming for comprehensive superiority across these metrics.

First, for the MR-CT Pelvis dataset, Table I shows that the proposed method achieves second-best performance in GC (0.43), NMI (0.41), and sharpness (99.8) while achieving the highest scores in KID (0.50). Although FID (16.2) ranks third, the overall performance across all metrics is notable for its comprehensive result. In contrast, the registration methods consistently show low GC scores compared to our method, with qualitative observations indicating structural misalignment with input (x) images, particularly in local areas (see Fig.3(d,e,g)). Moreover, as shown in Table I, registration methods generally yield lower sharpness values. On the other hand, I2I translation methods tend to show relatively inferior FID and KID scores, with qualitative results revealing issues such as hallucinations and incorrect intensity, indicating a failure to preserve the features of the reference (y) image (See Fig.4(d,e,f)).

Second, for the CBCT-CT Brain dataset, Table II indicates that the proposed method achieves top performance with GC (0.57), NMI (0.46), and FID (7.6), demonstrating comprehensive superiority across all metrics. By contrast, registration methods show relatively low GC and NMI scores, with qualitative evaluations further highlighting structural misalignment with input images (see Fig.3(k,l,n)). Notably, Table II shows consistently low sharpness values for registration methods, with the exception of GradICON (3D). On the other hand, I2I translation methods display inferior FID and KID scores in Table II, and qualitative observations reveal that the generated CT images do not preserve fine features of the cervical vertebrae as well as actual CT (reference images) (see the zoomed-in regions in Fig.4(j,k,l)). In contrast, the proposed method preserves fine features of the cervical vertebrae, showing the highest sharpness score and effectively maintaining these details, as shown in Fig.4(i).

2) Case study: Figure 5 presents a case study conducted on the MR-CT pelvis dataset to evaluate the proposed method's ability to preserve critical features of the actual CT (reference image). In clinical practice, CT urography is performed with contrast agents in patients with specific symptoms, and it is highly effective in diagnosing various urinary tract conditions, including hematuria and bladder cancer [72]–[74]. However, preserving these features is challenging for I2I translation, as these instances represent underrepresented classes within the dataset. This underrepresentation amplifies the ill-posed nature of the input-reference condition, leading to errors (see the red arrow in Fig.5(d)). In contrast, the proposed method not only preserves the contrast-enhanced features of the bladder in the actual CT (see the green arrow in Fig.5(c)), but also ensures structural alignment with the input MR. This capability suggests its potential clinical applicability in MR-CT-guided radiotherapy planning [75] for urinary tract-related diseases. Additionally, the proposed method preserves the features of the actual CT even in cases of an empty colon, which represents an underrepresented class, as shown in Fig.5(ii).

3) Downstream task analysis (Segmentation): To validate the practical utility of the proposed method, we assess segmentation performance as a downstream task on the MR-CT Pelvis dataset. In real radiotherapy planning, MR and CT images are aligned and mutually referenced [76]. In this experiment, we used TotalSegmentator [77] to segment major anatomical structures on MR and CT images. The segmentation was evaluated using the Dice coefficient and Hausdorff distance. For quantitative evaluation, manual refinement of the masks

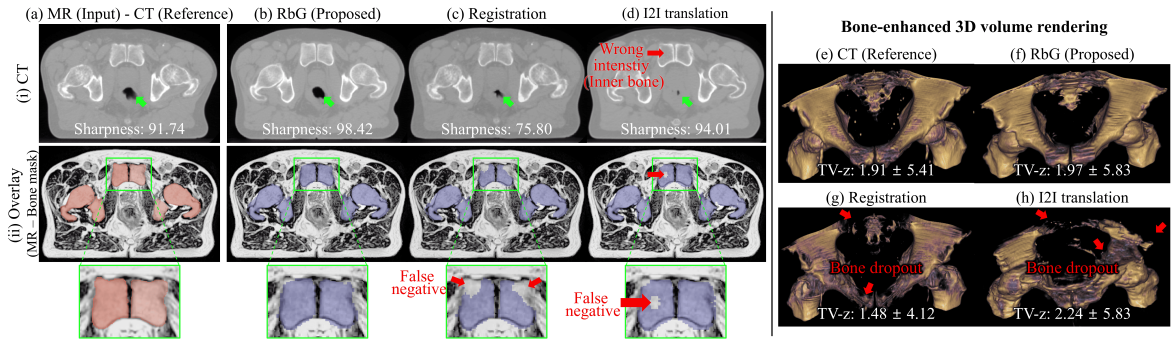

Fig. 6. Comparative analysis of downstream segmentation from 3D stacked CT images. Red regions show manually refined ground truth masks, while blue regions indicate masks generated by TotalSegmentator from CT images. Bone-enhanced 3D volume rendering results are shown in (e–f), with values indicating mean ± standard deviation of z-direction total variation (TV-z) between adjacent slices, reflecting inter-slice consistency. For comparison, Voxelmorph (2D) was used as a registration method, and MUNIT as an I2I translation method.

TABLE III
QUANTITATIVE COMPARISON OF DOWNSTREAM SEGMENTATION METRICS (DICE SCORE AND HAUSDORFF DISTANCE) BETWEEN 3D REGISTRATION NETWORKS (VOXELMORPH, LAPIRN) AND AN I2I TRANSLATION NETWORK (MUNIT). SIGNIFICANCE MARKERS ARE DEFINED IN TABLE I.

Metric: Dice Coefficient (↑)					
	CT(Ref)	Voxelmorph	LapIRN	MUNIT	Proposed
Hips (Bone)	0.71 ± 0.08 **	0.73 ± 0.02 **	0.69 ± 0.08 **	0.72 ± 0.05 **	0.75 ± 0.02
Femurs (Bone)	0.77 ± 0.14 *	0.81 ± 0.03 -	0.77 ± 0.16 **	0.79 ± 0.13 -	0.81 ± 0.03
Spinal cord	0.40 ± 0.29 **	0.46 ± 0.28 *	0.46 ± 0.26 *	0.16 ± 0.24 **	0.51 ± 0.18
Urinary bladder	0.55 ± 0.21 **	0.57 ± 0.23 **	0.51 ± 0.24 **	0.44 ± 0.33 **	0.66 ± 0.18
Metric: Hausdorff Distance (↓)					
	CT(Ref)	Voxelmorph	LapIRN	MUNIT	Proposed
Hips (Bone)	33.5 ± 23.2 **	11.7 ± 2.9 -	27.9 ± 16.8 **	25.4 ± 17.2 **	11.3 ± 1.9
Femurs (Bone)	14.8 ± 18.1 *	8.9 ± 1.0 **	12.6 ± 17.9 **	17.4 ± 19.6 **	7.7 ± 1.1
Spinal cord	12.2 ± 4.8 **	10.8 ± 2.8 **	10.1 ± 5.2 **	15.7 ± 10.9 **	8.4 ± 1.9
Urinary bladder	27.5 ± 36.0 **	22.7 ± 16.7 **	22.5 ± 11.7 **	36.3 ± 39.3 **	15.8 ± 9.1

was performed due to the inherent limitations of MR images in accurately representing hard tissues, which resulted in minor errors in bone masking. The output images, inferred in 2D axial slices, are stacked in 3D for each patient to enable segmentation processing.

Quantitatively, as shown in Table III, the proposed method achieves superior performance for hard tissues such as the hips and femurs. Additionally, for soft tissues like the spinal cord and urinary bladder, the proposed method records relatively high performance, though some lower scores are observed due to the inherent limitations of CT images in distinguishing soft tissue. Qualitatively, the proposed method's bone mask closely aligns with the bone region of MR images (see Fig. 6(ii-b)).

In contrast, the registration method shows false negatives in the mask with lower sharpness. While I2I methods demonstrate high sharpness, they fail to accurately capture the actual CT features inner part of the bone, resulting in false negatives in the mask (see the red arrow in Fig. 6(ii-d)). These observations emphasize that preserving the sharpness and intrinsic features of CT images is crucial for accurate segmentation [78], [79].

Additionally, to evaluate the anatomical fidelity and inter-slice consistency of the proposed method, the 2D outputs were stacked along the axial axis and visualized through bone-enhanced volume rendering using 3D Slicer [80]. The bone-enhanced renderings demonstrate that the proposed method effectively reconstructs bony structures comparable to the reference CT, indicating high anatomical fidelity (see Fig. 6(e,f)). Furthermore, to evaluate the z-direction inter-slice consistency despite the 2D nature of the proposed method, we computed the z-direction total variation (TV-z) between each slice k and its adjacent slice $k+1$, and then derived the mean and standard deviation of these values. The proposed method achieved TV-z statistics most closely matching those of the reference CT, indicating superior inter-slice consistency. Among the comparative methods, the registration approach exhibited lower TV-z values, mainly due to smoothing introduced during the registration process, while the image-to-image (I2I) translation method showed higher TV-z values, attributed to bone dropout and hallucination effects that caused discontinuities between slices. These results demonstrate that, although 2D-based, the proposed reference-augmented generation ensures inter-slice consistency and enhances overall generation stability.

B. Experiment on the Well-Aligned Dataset

1) Quantitative and qualitative evaluation: For retrospective validation, we use a well-aligned T1-T2 brain dataset and apply intentional misalignment to assess the misalignment robustness.

Quantitatively, as shown in Table IV, the proposed method demonstrates robust performance across various misalignment scenarios, outperforming conventional unsupervised I2I translation methods. Qualitatively, I2I translation methods (UNIT, MUNIT) show lower PSNR and SSIM scores and display hallucination artifacts (see the red arrow in Fig.7(d,e)). This highlights the effectiveness of the proposed method, which utilizes the reference image as guidance, in resolving the ambiguous relationship between the input and reference images.

An interesting observation from Table IV is that when rigid misalignment is applied, SSIM scores improve across all models, suggesting that slight uncertainty between paired data may contribute to model generalization. In contrast, non-rigid misalignment does not consistently lead to performance improvements.

TABLE IV

COMPARISON OF I2I TRANSLATION ON THE T1-T2 BRAIN DATASET WITH INTENTIONAL MISALIGNMENT TO ASSESS ROBUSTNESS. "MISALIGNED TEST SET" USES THE ALIGNED TEST SET AS GROUND TRUTH. SIGNIFICANCE MARKERS ARE DEFINED IN TABLE I.

Methods	PSNR ↑			SSIM ↑			LPIPS ↓ ($\times 10^{-1}$)		
	Aligned	Misaligned (Rigid)	Misaligned (Non-rigid)	Aligned	Misaligned (Rigid)	Misaligned (Non-rigid)	Aligned	Misaligned (Rigid)	Misaligned (Non-rigid)
Misaligned Test Set	–	23.7 ± 5.8	21.3 ± 2.1	–	0.78 ± 0.12	0.75 ± 0.06	–	0.14 ± 0.10	0.26 ± 0.21
CGAN	21.7 ± 1.4 **	22.5 ± 1.6 **	22.0 ± 1.2 **	0.76 ± 0.05 **	0.80 ± 0.05 **	0.75 ± 0.03 **	0.63 ± 0.28 **	0.66 ± 0.26 **	0.63 ± 0.26 *
SCGAN	21.3 ± 0.8 **	24.3 ± 2.3 **	21.8 ± 1.3 **	0.68 ± 0.05 **	0.85 ± 0.06 **	0.72 ± 0.05 **	0.81 ± 0.25 **	0.60 ± 0.22 *	0.61 ± 0.21 -
UNIT	22.5 ± 0.9 **	22.3 ± 1.0 **	21.7 ± 0.9 **	0.80 ± 0.05 **	0.81 ± 0.05 **	0.79 ± 0.05 **	0.80 ± 0.24 **	0.76 ± 0.21 **	0.91 ± 0.21 **
MUNIT	20.7 ± 1.3 **	21.1 ± 1.4 **	20.6 ± 1.3 **	0.78 ± 0.06 **	0.79 ± 0.04 **	0.79 ± 0.04 **	0.80 ± 0.27 **	0.87 ± 0.26 **	0.88 ± 0.25 **
Proposed	25.1 ± 1.7	25.4 ± 1.6	24.3 ± 1.3	0.85 ± 0.04	0.87 ± 0.04	0.84 ± 0.04	0.49 ± 0.16	0.47 ± 0.17	0.54 ± 0.16

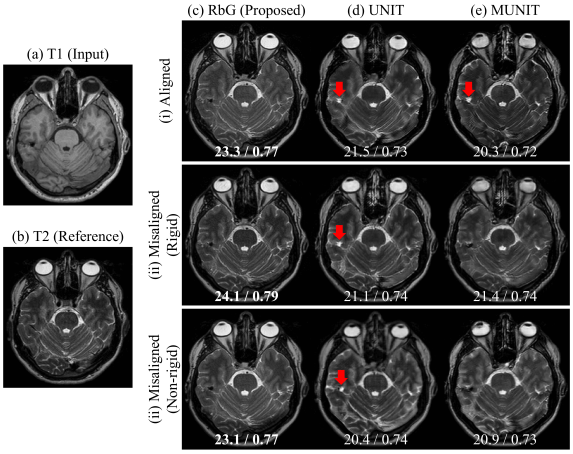

Fig. 7. Comparison of I2I translation (UNIT, MUNIT) on the well-aligned T1-T2 Brain dataset. The red arrow denotes the Hallucination artifact. The numbers superimposed on each image represent the PSNR and SSIM values.

TABLE V

ABLATION STUDY (DOWNSAMPLE FACTORS IN STAGE 1). THE FACTOR (H,W) REPRESENTS GLOBAL FEATURE TRANSFER (ORIGINAL ADAIN), WHILE SMALLER FACTORS ENABLE LOCALIZED TRANSFER. THE ASTERISK (*) INDICATES THE PROPOSED METHOD'S CHOICE.

No.	Downsample factor	GC ↑	NMI ↑	FID ↓
(a)	(H, W)	0.42 ± 0.05	0.41 ± 0.04	25.7
(b)	64	0.43 ± 0.05	0.41 ± 0.04	28.6
(c)*	32	0.51 ± 0.08	0.43 ± 0.04	25.1
(d)	16	0.50 ± 0.06	0.44 ± 0.04	26.8
(e)	8	0.41 ± 0.06	0.41 ± 0.04	15.7

TABLE VI

ABLATION STUDY (CORE COMPONENTS OF DACA BLOCK IN STAGE 2). "PS" DENOTES THE LOCAL PATCH SIZE IN CROSS ATTENTION. "DF" INDICATES WHETHER THE DEFORMATION FIELD IS USED.

No.	PS	DF	GC ↑	NMI ↑	FID ↓
(a)	1	✓	0.28	0.36	2.0
(b)	20	✓	0.39	0.39	17.1
(c)	28		0.41	0.37	49.6
(d)*	28	✓	0.43	0.41	16.2

C. Ablation Study

We conduct the ablation study on the MR-CT pelvis dataset, as it presents challenging conditions due to significant modality discrepancy in the MR-CT pair and complex misalignment in the pelvic region.

1) Downsample factor in Stage 1: In Stage 1, we compare performance based on the "downsample factor," a key parameter of the proposed PAdaIN method. A larger downsample factor results in more global feature transfer, while a smaller factor allows for more localized and detailed feature transfer. As shown in Table V, the selected factor of 32 demonstrates comprehensive superiority across all metrics. In contrast, a factor of 8 achieves the highest FID but relatively lower GC. Additionally, a factor of (H, W), equivalent to the existing AdaIN [32] method for global feature transfer, shows inferior overall performance compared to the proposed method. The results of factor (H, W) and factor 32 are qualitatively compared in Fig.8(i-c,d), where the proposed Patch Adaptive Instance Normalization (PAdaIN) corresponding to factor 32 is observed to be more effective in preserving the features of the CT image. The optimal factor depends on the dataset; severe misalignment benefits from larger factors for stable global transfer, but excessively large values may degrade local detail preservation, as shown in Table V-(b).

2) Core components of DACA block in Stage 2: We evaluated the effectiveness of two key components in the proposed Stage 2 DACA block: the local patch size (PS) in cross-attention and the use of deformation fields (DF) (see Fig.1(c)). Without DF, standard cross-attention is applied. As shown in Table VI, when PS is set to 1 with DF, FID scores are notably high, but GC and NMI scores are considerably low. This outcome suggests that solely using the deformation field fails to match perfectly aligned pixels, resulting in near-replication of the reference image. Conversely, when PS is set to 28 without DF, performance declines relative to the proposed method, indicating that cross-attention between misaligned features can introduce errors. These results highlight the effectiveness of setting an appropriate PS and incorporating DF to generate high-quality output images.

3) Combination of loss function in Stage 2: In Stage 2, two different types of loss functions are combined for training: one is the loss between the input (x) image, and the other is the loss between the reference (y) image. For the loss between the input (x) image, we compared contrastive loss ($\mathcal{L}_{\text{patchNCE}}$) with modality-independent neighborhood descriptor (MIND) loss ($\mathcal{L}_{\text{mind}}$) [20], commonly used to ensure structural alignment across modalities. For the loss between the reference (y) image, we compared conventional L1 loss ($\mathcal{L}_1$) with contextual loss ($\mathcal{L}_{\text{cx}}$).

As shown in Table VII, the combination of $\mathcal{L}_{\text{patchNCE}}$ and

TABLE VII

ABLATION STUDY (COMBINATION OF LOSS FUNCTION IN STAGE 2).

No.	with input		with reference		GC ↑	NMI ↑	FID ↓
	$\mathcal{L}_{\text{mind}}$	$\mathcal{L}_{\text{patchNCE}}$	$\mathcal{L}_1$	$\mathcal{L}_{\text{cx}}$			
(a)	✓			✓	0.40 ± 0.07	0.39 ± 0.04	19.4
(b)		✓	✓		0.45 ± 0.05	0.40 ± 0.04	78.1
(c)	✓		✓		0.48 ± 0.07	0.45 ± 0.04	34.3
(d)*		✓		✓	0.43 ± 0.07	0.41 ± 0.04	16.2

TABLE VIII
DATA EFFICIENCY ANALYSIS. COMPARISON OF PERFORMANCE BASED ON TRAINING SET PROPORTION. SIGNIFICANCE MARKERS ARE DEFINED IN TABLE I.

Train Set	Method	GC↑	NMI↑	FID↓
10%	UNIT	0.36 ± 0.03 **	0.39 ± 0.04 **	52.1
	MUNIT	0.30 ± 0.04 **	0.34 ± 0.03 **	69.4
	RbG (Proposed)	0.46 ± 0.06 **	0.40 ± 0.04 **	30.0
20%	UNIT	0.45 ± 0.06 **	0.37 ± 0.04 **	43.3
	MUNIT	0.38 ± 0.04 **	0.38 ± 0.04 **	52.3
	RbG (Proposed)	0.46 ± 0.06 **	0.40 ± 0.04 **	30.0
100%	UNIT	0.45 ± 0.06 -	0.36 ± 0.03 **	32.2
	MUNIT	0.34 ± 0.06 **	0.33 ± 0.03 **	28.9
	RbG (Proposed)	0.43 ± 0.07	0.41 ± 0.04	16.2

$\mathcal{L}_{\text{cx}}$, selected for the proposed method, achieved the highest FID score. Conversely, using $\mathcal{L}_1$ loss resulted in high GC scores but significantly lower FID scores. This suggests that $\mathcal{L}_{\text{cx}}$ is more advantageous for the effective preserving of misaligned features. The qualitative results for each combination of loss functions are shown in Fig.8(ii), where the proposed combination of $\mathcal{L}_{\text{patchNCE}}$ and $\mathcal{L}_{\text{cx}}$ demonstrates the best preservation of the bone structures in the generated CT.

4) Data efficiency analysis: To assess the data efficiency of the proposed method, we gradually reduced the proportion of the training set and compared performance. We simulated limited data training by randomly selecting 20% and 10% of the training set. Quantitatively, as shown in Table VIII, the proposed method exhibited minimal performance fluctuations in GC and NMI in different proportions of the training set. While FID performance deteriorated, the proposed method still consistently outperformed other methods.

In contrast, for I2I translation methods (UNIT, MUNIT), GC and NMI unexpectedly increased when the training set was reduced from 100% to 20%, while FID experienced a significant degradation. This suggests that these methods may overfit smaller datasets, maintaining structural alignment with the input (x) image but failing to capture features of the reference (y) image accurately. This experiment demonstrates that the proposed method, which uses self-supervised learning with the reference image as additional guidance, exhibits relatively robust performance even with limited datasets.

VI. CONCLUSION

In this paper, we present the Register by Generation (RbG) framework, a novel approach for cross-modality image alignment that uses a generation-based strategy. By integrating a reference-augmented synthesis and refinement process, RbG achieves precise alignment across modalities while preserving the unique contrast and texture of reference images. Our two-stage approach combines the Patch Adaptive Instance Normalization (PAdaIN) and Deformation-Aware Cross-Attention (DACA) blocks to perform accurate alignment and feature fusion, addressing challenges in complex anatomical deformations and modality-specific differences.

The RbG framework has been rigorously validated on multiple datasets. Our method has consistently shown strong structural alignment and effective feature preservation, outperforming conventional registration and I2I translation approaches, particularly in handling local anatomical deformations. The robustness of our approach was further validated through

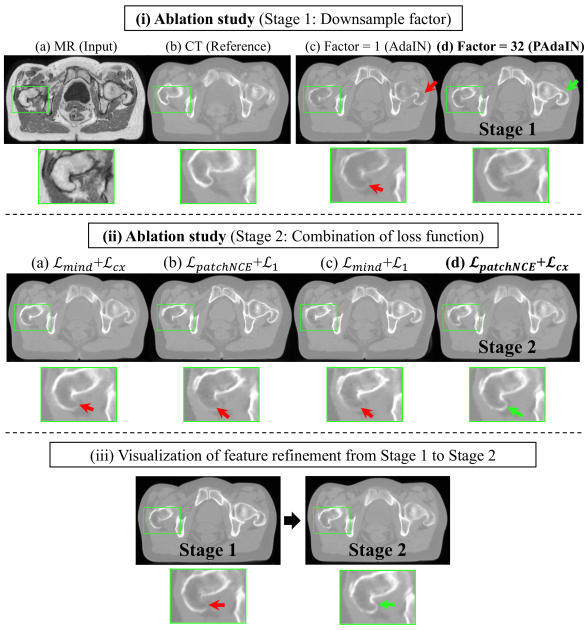

Fig. 8. Ablation study visualization. (i) Comparison of downsampling factors in Stage 1 (Table IV). (ii) Comparison of loss function combinations in Stage 2 (Table VI). (iii) Stage 1 output refined by Stage 2.

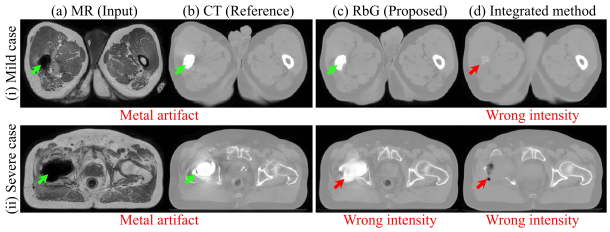

Fig. 9. Failure cases with metal artifacts. Compared to the integrated method (RegGAN), the proposed method preserves features in mild cases (i) but struggles in severe cases (ii). RegGAN fails in both scenarios.

case studies, downstream segmentation tasks, and retrospective validations, where the proposed framework maintained high-quality feature preservation.

Certain limitations remain in the current scope of this study. First, failure cases remain, particularly in the presence of metal artifacts in MR-CT pairs (e.g., from implants). These scenarios present significant difficulties due to severe signal loss in MRI around metal objects, which degrades the performance of Stage 1 and complicates the recovery of CT features in Stage 2 (see Fig. 9).

Second, the current framework processes 2D slices, which are then stacked to produce a 3D result. Expanding to a full 3D approach is currently constrained by GPU memory limitations. However, future work will aim to address this by utilizing memory-efficient transformer architectures and other optimized algorithms. We anticipate that this enhancement will further improve alignment accuracy. Notably, as our current 2D approach already surpasses existing 3D registration methods (as demonstrated in Tables I and II) and shows relatively high inter-slice consistency (Fig. 6(e–h)), extending to 3D is expected to yield even greater performance improvements.

VII. References

[1] M. Soltaninejad, G. Yang, T. Lambrou, N. Allinson, T. L. Jones, T. R. Barrick, F. A. Howe, and X. Ye, "Supervised learning based multimodal mri brain tumour segmentation using texture features from supervoxels," *Computer methods and programs in biomedicine*, vol. 157, pp. 69–84, 2018. 1

[2] R. Han, C. K. Jones, J. Lee, P. Wu, P. Vagdargi, A. Uneri, P. A. Helm, M. Luciano, W. S. Anderson, and J. H. Siewerdsen, "Deformable mr-ct image registration using an unsupervised, dual-channel network for neurosurgical guidance," *Medical image analysis*, vol. 75, p. 102292, 2022. 1

[3] L. Bi, M. Fulham, N. Li, Q. Liu, S. Song, D. D. Feng, and J. Kim, "Recurrent feature fusion learning for multi-modality pet-ct tumor segmentation," *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106043, 2021. 1

[4] F. Wang, C. Cheng, W. Cao, Z. Wu, H. Wang, W. Wei, Z. Yan, and Z. Liu, "Mfcnet: A multi-modal fusion and calibration networks for 3d pancreas tumor segmentation on pet-ct images," *Computers in Biology and Medicine*, vol. 155, p. 106657, 2023. 1

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 1

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762 1

[7] B. Kim, Y. Zhuang, T. S. Mathai, and R. M. Summers, "Otmorph: Unsupervised multi-domain abdominal medical image registration using neural optimal transport," *IEEE Transactions on Medical Imaging*, 2024. 1, 3

[8] Y. Huang, L. Shao, and A. F. Frangi, "Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 815–827, 2017. 1

[9] O. Dalmaz, M. Yurt, and T. Çukur, "Resvit: residual vision transformers for multimodal medical image synthesis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022. 1

[10] F. Darzi and T. Bocklitz, "A review of medical image registration for different modalities," *Bioengineering*, vol. 11, no. 8, p. 786, 2024. 1

[11] T. Han, J. Wu, W. Luo, H. Wang, Z. Jin, and L. Qu, "Review of generative adversarial networks in mono-and cross-modal biomedical image registration," *Frontiers in Neuroinformatics*, vol. 16, p. 933230, 2022. 1

[12] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1153–1190, 2013. 1

[13] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1788–1800, 2019. 2, 3, 5, 7

[14] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 2

[15] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023. 2

[16] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024. 2

[17] S. Sun, K. Han, D. Kong, H. Tang, X. Yan, and X. Xie, "Topology-preserving shape reconstruction and registration via neural diffeomorphic flow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 845–20 855. 2

[18] S. Liu, B. Yang, Y. Wang, J. Tian, L. Yin, and W. Zheng, "2d/3d multimode medical image registration based on normalized cross-correlation," *Applied Sciences*, vol. 12, no. 6, p. 2828, 2022. 2

[19] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997. 2, 3

[20] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel, "Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical image analysis*, vol. 16, no. 7, pp. 1423–1435, 2012. 2, 3, 11

[21] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen, "Unsupervised deformable registration for multi-modal images via disentangled representations," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 249–261. 2

[22] H. R. Boveiri, R. Khayami, R. Javidan, and A. Mehdizadeh, "Medical image registration using deep neural networks: a comprehensive review," *Computers & Electrical Engineering*, vol. 87, p. 106767, 2020. 2

[23] F. S. Zadeh, S. Molani, M. Orouskhani, M. Rezaei, M. Shafiei, and H. Abbasi, "Generative adversarial networks for brain images synthesis: A review," *arXiv preprint arXiv:2305.15421*, 2023. 2

[24] M. Boulanger, J.-C. Nunes, H. Chourak, A. Largent, S. Tahri, O. Acosta, R. De Crevoisier, C. Lafond, and A. Barateau, "Deep learning methods to generate synthetic ct from mri in radiotherapy: A literature review," *Physica Medica*, vol. 89, pp. 265–281, 2021. 2

[25] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, "Deep learning based synthetic-ct generation in radiotherapy and pet: a review," *Medical physics*, vol. 48, no. 11, pp. 6537–6566, 2021. 2

[26] Y. Ge, D. Wei, Z. Xue, Q. Wang, X. Zhou, Y. Zhan, and S. Liao, "Unpaired mr to ct synthesis with explicit structural constrained adversarial learning," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1096–1099. 2, 3

[27] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato, "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. 2, 3

[28] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, J. L. Prince, and Z. Xu, "Unsupervised mr-to-ct synthesis using structure-constrained cyclegan," *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 4249–4261, 2020. 2, 3, 7

[29] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, Ş. Özturk, A. Güngör, and T. Çukur, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Transactions on Medical Imaging*, 2023. 2, 3

[30] J. P. Cohen, M. Luck, and S. Honari, "Distribution matching losses can hallucinate features in medical image translation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018, pp. 529–536. 2, 3

[31] S. Kim, C. Jin, T. Diethe, M. Figini, H. F. Tregidgo, A. Mullokandov, P. Teare, and D. C. Alexander, "Tackling structural hallucination in image translation with local diffusion," *arXiv preprint arXiv:2404.05980*, 2024. 2, 3

[32] X. Huang and S. 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. 2, 4, 11

[33] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337–2346. 2, 4

[34] C. Hémon, V. Boussot, B. Texier, J.-L. Dillenseger, and J.-C. Nunes, "Guiding unsupervised cbct-to-ct synthesis using content and style representation by an enhanced perceptual synthesis (creps) loss," in *SynthRAD2023 Challenge, MICCAI 2023,*, 2023. 2

[35] L. Kong, C. Lian, D. Huang, Y. Hu, Q. Zhou *et al.*, "Breaking the dilemma of medical image-to-image translation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1964–1978, 2021. 2, 3, 7

[36] Y. Liu, W. Wang, Y. Li, H. Lai, S. Huang, and X. Yang, "Geometry-consistent adversarial registration model for unsupervised multi-modal medical image registration," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3455–3466, 2023. 3

[37] Z. Ghassabi, J. Shanbehzadeh, A. Sedaghat, and E. Fatemizadeh, "An efficient approach for robust multimodal retinal image registration based on ur-sift features and piifd descriptors," *EURASIP Journal on Image and Video Processing*, vol. 2013, pp. 1–16, 2013. 3

[38] G. Song, J. Han, Y. Zhao, Z. Wang, and H. Du, "A review on medical image registration as an optimization problem," *Current Medical Imaging*, vol. 13, no. 3, pp. 274–283, 2017. 3

[39] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du, "Transmorph: Transformer for unsupervised medical image registration," *Medical image analysis*, vol. 82, p. 102615, 2022. 3, 7

[40] L. Tian, H. Greer, F.-X. Vialard, R. Kwitt, R. S. J. Estépar, R. J. Rushmore, N. Makris, S. Bouix, and M. Niethammer, "Gradicon: Approximate diffeomorphisms via gradient inverse consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 084–18 094. 3, 7

[41] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712–721, 1999. 3

[42] Y. Zhao, X. Chen, B. McDonald, C. Yu, A. S. Mohamed, C. D. Fuller, L. E. Court, T. Pan, H. Wang, X. Wang *et al.*, "A transformer-based hierarchical registration framework for multimodality deformable image registration," *Computerized Medical Imaging and Graphics*, vol. 108, p. 102286, 2023. 3

[43] X. Song, H. Chao, X. Xu, H. Guo, S. Xu, B. Turkbey, B. J. Wood, T. Sanford, G. Wang, and P. Yan, "Cross-modal attention for multi-modal image registration," *Medical Image Analysis*, vol. 82, p. 102612, 2022. 3

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232. 3, 7

[45] R. Wang, A. F. Heimann, M. Tannast, and G. Zheng, "Cyclesgan: A cycle-consistent and semantics-preserving generative adversarial network for unpaired mr-to-ct image synthesis," *Computerized Medical Imaging and Graphics*, vol. 117, p. 102431, 2024. 3

[46] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in neural information processing systems*, vol. 30, 2017. 3, 7

[47] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189. 3, 7

[48] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51. 3

[49] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangled representation learning in cardiac image analysis," *Medical image analysis*, vol. 58, p. 101535, 2019. 3

[50] S. Reaungamornrat, H. Sari, C. Catana, and A. Kamen, "Multimodal image synthesis based on disentanglement representations of anatomical and modality specific features, learned using uncooperative relativistic gan," *Medical image analysis*, vol. 80, p. 102514, 2022. 3

[51] Y. Zhang, C. Li, Z. Dai, L. Zhong, X. Wang, and W. Yang, "Breath-hold cbct-guided cbct-to-ct synthesis via multimodal unsupervised representation disentanglement learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2313–2324, 2023. 3

[52] C. Hemon, B. Texier, H. Chourak, A. Simon, I. Bessières, R. de Crevoisier, J. Castelli, C. Lafond, A. Barateau, and J.-C. Nunes, "Indirect deformable image registration using synthetic image generated by unsupervised deep learning," *Image and Vision Computing*, vol. 148, p. 105143, 2024. 3

[53] A. Korotin, D. Selikhanovych, and E. Burnaev, "Neural optimal transport," *arXiv preprint arXiv:2201.12220*, 2022. 3

[54] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "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. 3

[55] J. Cao, J. Liang, K. Zhang, Y. Li, Y. Zhang, W. Wang, and L. V. Gool, "Reference-based image super-resolution with deformable attention transformer," in *European conference on computer vision*. Springer, 2022, pp. 325–342. 3

[56] L. Xiang, Y. Chen, W. Chang, Y. Zhan, W. Lin, Q. Wang, and D. Shen, "Ultra-fast t2-weighted mr reconstruction using complementary t1-weighted information," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*. Springer, 2018, pp. 215–223. 3

[57] S. U. Dar, M. Yurt, M. Shahdloo, M. E. Ildız, B. Tınaz, and T. Çukur, "Prior-guided image reconstruction for accelerated multi-contrast mri via generative adversarial networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1072–1087, 2020. 3

[58] K. Xuan, L. Xiang, X. Huang, L. Zhang, S. Liao, D. Shen, and Q. Wang, "Multimodal mri reconstruction assisted with spatial alignment network," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2499–2509, 2022. 3

[59] P. Guo and V. M. Patel, "Reference-based mri reconstruction using texture transformer," in *Medical Imaging with Deep Learning*, 2023. 3

[60] K. H. Kim, W.-J. Do, and S.-H. Park, "Improving resolution of mr images with an adversarial network incorporating images with different contrast," *Medical physics*, vol. 45, no. 7, pp. 3120–3131, 2018. 4

[61] R. Souza, Y. Beauferris, W. Loos, R. M. Lebel, and R. Frayne, "Enhanced deep-learning-based magnetic resonance image reconstruction by leveraging prior subject-specific brain imaging: Proof-of-concept using a cohort of presumed normal subjects," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1126–1136, 2020. 4

[62] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410. 4

[63] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711. 6

[64] R. Mechrez, I. Talmi, and L. 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. 6

[65] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014. 6

[66] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021. 6

[67] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE transactions on medical imaging*, vol. 19, no. 2, pp. 143–150, 2000. 6

[68] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008. 7

[69] T. C. Mok and A. C. 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. 7

[70] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, "Unsupervised multi-modal image registration via geometry preserving image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 410–13 419. 7

[71] R. F. Woolson, "Wilcoxon signed-rank test," *Encyclopedia of biostatistics*, vol. 8, 2005. 8

[72] J. Zhang, S. Gerst, R. A. Lefkowitz, and A. Bach, "Imaging of bladder cancer," *Radiologic Clinics*, vol. 45, no. 1, pp. 183–205, 2007. 9

[73] A. Kawashima, T. J. Vrtiska, A. J. LeRoy, R. P. Hartman, C. H. McCollough, and B. F. King Jr, "Ct urography," *Radiographics*, vol. 24, no. suppl_1, pp. S35–S54, 2004. 9

[74] N. C. Cowan, "Ct urography for hematuria," *Nature Reviews Urology*, vol. 9, no. 4, pp. 218–226, 2012. 9

[75] J. Brunt, "Computed tomography–magnetic resonance image registration in radiotherapy treatment planning," *Clinical oncology*, vol. 22, no. 8, pp. 688–697, 2010. 9

[76] H. Chandarana, H. Wang, R. Tijssen, and I. J. Das, "Emerging role of mri in radiation therapy," *Journal of Magnetic Resonance Imaging*, vol. 48, no. 6, pp. 1468–1478, 2018. 9

[77] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang *et al.*, "Totalsegmentator: robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, 2023. 9

[78] H. Zunair and A. B. Hamza, "Sharp u-net: Depthwise convolutional network for biomedical image segmentation," *Computers in biology and medicine*, vol. 136, p. 104699, 2021. 10

[79] Y. Choi, M. A. Al-Masni, K.-J. Jung, R.-E. Yoo, S.-Y. Lee, and D.-H. Kim, "A single stage knowledge distillation network for brain tumor segmentation on limited mr image modalities," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107644, 2023. 10

[80] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, "3d slicer: a platform for subject-specific image analysis, visualization, and clinical support," in *Intraoperative imaging and image-guided therapy*. Springer, 2013, pp. 277–289. 10