Unsupervised Exemplar-Based Image-to-Image Translation and Cascaded Vision Transformers for Tagged and Untagged Cardiac Cine MRI Registration

Meng Ye[1], Mikael Kanski[2], Dong Yang[3], Leon Axel[2], Dimitris Metaxas[1]
[1]Rutgers University, [2]New York University School of Medicine, [3]NVIDIA
{my389, dnm}@cs.rutgers.edu

Abstract

Multi-modal registration between tagged and untagged cardiac cine magnetic resonance (MR) images remains difficult, due to the domain gap and large deformations between the two modalities. Recent work using an image-to-image translation (I2I) module to overcome the domain gap can convert the multi-modal into a mono-modal registration task and take advantage of advanced mono-modal registration architectures. However, they often ignore two issues: the sample-specific style of each image to be registered during I2I and large hybrid rigid and non-rigid deformations between modalities. We first propose an exemplar-based I2I module capable of unsupervised cross-domain correspondence learning to enforce the style consistency between the fake image and the image to be registered. Then we propose an efficient cascaded vision transformer-based registration network to predict both the affine and non-rigid deformations, in which a single feature embedding subnetwork is shared by the two stages of deformation prediction. We validated our method on a clinical cardiac MR dataset with paired but unaligned untagged and tagged MR images. The results show that our method outperforms traditional methods significantly in terms of the I2I quality and multi-modal image registration accuracy.

1. Introduction

Multi-modal medical imaging provides complementary information for clinical disease diagnosis. As shown in Fig. 1, for dynamic cardiac magnetic resonance (MR) imaging, we have two different 2D cine imaging modalities: traditional untagged cine MR (cMR) and tagged cine MR (tMR) imaging [1]. While cMR provides the gold standard imaging modality for global cardiac function evaluation, tMR is the gold standard for regional myocardium (Myo) wall motion quantification and strain estimation [3]. To extract the region-of-interest (ROI) for the Myo wall, we need to segment such ROIs from images. However, due to the tagged blood in early time frames, which obscures the boundary between the Myo wall and blood pool, segmenting the Myo region on tMR frames remains difficult [33]. Although recent data-driven models, such as U-Net [36] and V-Net [28], advance the segmentation performance on medical images, training such models needs a large dataset and annotations. Several large annotated cMR datasets, such as UK Biobank [32] and ACDC [5] make it possible to train a segmentation model to predict the Myo masks on cMR data. However, tMR data is less common, compared with cMR, as are annotations on tMR images. In clinical applications, tMR scans usually follow cMR scans, resulting in paired cMR and tMR imaging data for each patient. Robust multi-modal image registration between tMR and cMR thus makes segmenting the Myo wall on tMR frames feasible, by propagating the Myo masks from the cMR images to the tMR images with the associated deformation fields.

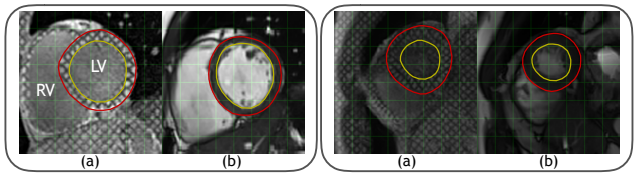

Figure 1. Two image pairs of (a) tagged cine MR (tMR) and (b) untagged cine MR (cMR). Note the diverse sample-specific styles within each modality and large inter-modality deformations for each paired data. Red and yellow contours show the epi- and endo-cardial borders of the left ventricle myocardium (Myo) wall. Green grids are for alignment visual inspection. LV: left ventricle; RV: right ventricle.

To register a cMR image to a tMR image means to warp the shape or **content** of specific ROIs from the **moving** image (cMR) to the **fixed** image (tMR) according to some spatial mapping function. The task is challenging, due to the domain gap and potential large deformations between the two different imaging modalities. One can easily observe from Fig. 1 that the appearance or **style** of the two images is distinct from each other, because they are produced by dif-

ferent MR imaging sequences. Traditional multi-modality image registration methods rely on developing robust similarity metrics, e.g., MIND features [14], to define an energy function for the registration model. Recent deep learning-based models either seek similarity metrics in a common content feature space [34, 41] or use a generator to translate the image style from the **source** domain to the **target** domain and thus convert multi-modal registration into a mono-modal registration [21, 42]. For the latter approach, there are two aspects that determine the success: content preservation and sample-specific style coherence. Recent work, however, often ignores the latter aspect, which degrades the registration performance. As the two samples shown in Fig. 1 demonstrate, for each modality, different samples may manifest specific styles, e.g., both the tMR and cMR images of the left sample show a brighter style than those of the right one. Our method falls into the multi-to-mono modality transform category. Different from previous methods, we aim to learn sample-specific styles during image-to-image translation for multi-modal registration and propose to use the target domain image as the style exemplar to guide the generator in an **unsupervised** fashion.

The large deformation challenge between different modalities is due to imaging condition changes during separate scans. As shown in Fig. 1, change of the breath-holding location and cardiac deformation can result in large and hybrid rigid and non-rigid motion, even between the paired tMR and cMR data. To deal with such large deformations, we design a cascaded vision transformer-based registration network to predict both affine and non-rigid deformations simultaneously and efficiently.

Our contributions in this work can be summarized as follows: (1) We propose a novel unsupervised multi-modal medical image registration method, which achieves a high registration accuracy with efficient inference. (2) We propose an unsupervised exemplar-based image-to-image translation module, which can efficiently learn the cross-domain correspondence without having strictly aligned training data pairs and significantly enhance the multi-modal image registration performance. (3) We propose an efficient cascaded vision transformer-based registration module, via the design of a shared subnetwork for different stages of deformation estimation, which can predict the large and hybrid affine and non-rigid deformations between modalities accurately.

2. Related Work

2.1. Image-to-Image Translation

While there exist multi-domain image-to-image translation (I2I) tasks [15, 20, 55], in this work, we focus on two-domain I2I tasks [17, 22, 54]. Given a source domain A and a target domain B, the goal of the I2I task is to transfer the style of a target domain B to the source domain A, while keeping the content of the input source domain image $\boldsymbol{x} \in A$ invariant. Most prominent data-driven learning-based methods rely on the use of generative adversarial networks (GANs) [13]. These methods aim to train a mapping network, i.e., the generator G, to generate a fake image $\hat{\boldsymbol{y}}$ from the input source domain image $\boldsymbol{x} \in A$ which makes the discriminator D fail to distinguish it from the target domain image $\boldsymbol{y} \in B$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are paired and aligned, it is a supervised I2I task. However, it is difficult to obtain paired and aligned training data, especially for the medical imaging domain, so recent efforts have focused more on unsupervised I2I tasks.

Two kinds of losses are designed to achieve content-preserving in I2I tasks, i.e., cycle-consistency loss used in a two-sided architecture [15,20,22,54] and other feature-level losses used in a one-sided architecture [18, 27, 31, 39, 50]. In general, style transfer can be achieved by using an adversarial loss which makes the style of fake images indistinguishable from that of the real ones. However, the adversarial loss only makes the generator learn the averaged style distribution of the target domain. To learn a finer target style, exemplar-based I2I frameworks have been introduced in [23, 25, 44–46, 51, 53]. However, these works require strictly aligned training data $\boldsymbol{x}$ and $\boldsymbol{y}$ to learn the correspondence between source image $\boldsymbol{x}$ and the exemplar $\boldsymbol{z}$, which are supervised I2I methods. We inherit the idea of using an exemplar image to guide the generator learning fine style of each specific image sample to be registered, but drive the cross-domain correspondence learning process in an unsupervised fashion.

2.2. Multi-Modality Medical Image Registration

Image registration aims to find the spatial mapping of corresponding contents in an image pair. Multi-modal registration is more difficult than a mono-modal task because of potentially severe intensity distortions and large deformations between modalities. Previous methods focus on developing robust similarity metrics to intensity distortions. Mutual information has been successfully used in rigid multi-modal registration [26, 40]. A recent work measures the mutual information through a jointly learned multi-scale and multi-modality embedding space for non-rigid image registration [11]. The modality-independent neighborhood descriptor (MIND) [14] is an image structural, instead of intensity, representation and is robust to intensity distortions, but it is computationally expensive. Some recent works construct the modality-independent similarity metric in a deep content feature space with the use of a pre-trained content encoder [34, 41]. Some efforts have been made to convert the multi-modal to mono-modal registration by using an I2I network [21,42]. However, these works ignore the sample-specific style during I2I. Large defor-

mations between modalities consist of rigid and non-rigid motions. Although convolutional neural networks (CNNs) could be trained to predict the global affine [49] and local non-rigid deformations [10], they fail to capture long-range dependencies of image features, due to the weighting sharing and locality inductive biases, resulting in sub-optimal registration performance [38]. Vision transformers (ViT) using self-attention mechanism to model long-range dependencies have been introduced to image classification tasks and dense prediction tasks, such as image segmentation and registration [12, 24, 52]. Although improved affine or deformable registration performance has been achieved by novel ViTs [6, 7, 30, 47], predicting hybrid affine and non-rigid deformations simultaneously and efficiently with ViTs still needs exploration. We aim to design a novel cascaded ViT-based registration network to predict large deformations between modalities accurately.

3. Methodology

Our novel multi-modal image registration method consists of an unsupervised `exemplar`-based I2I and `cascaded` vision `transformers` (`ECaT`). Although our method could be easily extended to other multi-modal image registration problems, without loss of generality, we focus on the challenging task for cMR registration with tMR. As shown in Fig. 2 (a), we have two modules in the pipeline: a style reference-augmented I2I network, and a cascaded affine and non-rigid registration network. Below, we detail each module.

3.1. Unsupervised Cross-Domain Correspondence Learning for Style Reference-Augmented I2I

We use an examplar-based I2I network to translate the tMR image ($\boldsymbol{x}$) to a fake cMR image ($\hat{\boldsymbol{y}}$), which serves as the fixed image for the downstream registration task. We input the real image $\boldsymbol{y}$ as the style reference (SR) into the generator G to learn a specific style from each individual $\boldsymbol{y}$. As shown in Fig. 2 (b), we divide G into an encoder G_e and a decoder G_d. G_e is shared for extracting features from $\boldsymbol{x}$ and $\boldsymbol{y}$: $\boldsymbol{F}_x = G_e(\boldsymbol{x})$, $\boldsymbol{F}_y = G_e(\boldsymbol{y})$, where $\boldsymbol{F}_x, \boldsymbol{F}_y \in \mathbb{R}^{H\times W\times C}$. With the feature representations, cross-domain correspondence is built with a correlation matrix $\mathcal{M} \in \mathbb{R}^{HW\times HW}$ [45], each entry of which is defined by the similarity of $\boldsymbol{F}_x$ at location i and $\boldsymbol{F}_y$ at location j:

$$\mathcal{M}(i,j) = \frac{(\boldsymbol{F}_x(i) - \boldsymbol{\mu}_{F_x}) \cdot (\boldsymbol{F}_y(j) - \boldsymbol{\mu}_{F_y})}{\left\|\boldsymbol{F}_x(i) - \boldsymbol{\mu}_{F_x}\right\|_2 \left\|\boldsymbol{F}_y(j) - \boldsymbol{\mu}_{F_y}\right\|_2}, \quad (1)$$

where $\boldsymbol{\mu}_{F_x}$ and $\boldsymbol{\mu}_{F_y}$ are the mean feature vectors, $\boldsymbol{F}_x, \boldsymbol{F}_y \in \mathbb{R}^{HW\times C}$ are reshaped vectors. We then align the features $\boldsymbol{F}_y$ with $\boldsymbol{F}_x$ by collecting the most correlated pixels in $\boldsymbol{F}_y$ and calculating the weighted average by $\mathcal{M}$:

$$\boldsymbol{F}_{y\to x}(i) = \sum_j \underset{j}{softmax}(\mathcal{M}(i,j)/\tau) \cdot \boldsymbol{F}_y(j), \quad (2)$$

where τ controls the sharpness of softmax and we empirically set it as $5e^{-3}$. Then we reshape $\boldsymbol{F}_x$ and $\boldsymbol{F}_{y\to x}$ as $H \times W \times C$ and concatenate them together as the input to the decoder G_d, which gives the translated fake image: $\hat{\boldsymbol{y}} = G_d(\boldsymbol{F}_x, \boldsymbol{F}_{y\to x})$.

Previous methods [45, 46] used strictly aligned paired data $\{\boldsymbol{x}, \boldsymbol{y}\}$ to train the cross-domain correspondence learning process in a supervised way. However, the medical imaging domain usually lacks such aligned data. We thus introduce a content-preserving loss to learn the cross-domain correspondence, along with image translation, in an unsupervised way. Here, while the content-preserving loss explicitly regularizes the synthesised image $\hat{\boldsymbol{y}}$ to avoid content distortion from the source image $\boldsymbol{x}$, it also implicitly regularizes a plausible cross-domain correspondence between the style exemplar $\boldsymbol{y}$ and $\boldsymbol{x}$. We show an example in Fig. 5. While our core idea is the unsupervised learning of sample-specific style transferring to benefit the downstream mono-modal registration, we do not focus on the designing of efficient content-preserving losses. We use a self-similarity based one proposed in the work [50]. We first extract the features $\boldsymbol{c}_x$ and $\boldsymbol{c}_{\hat{y}}$ from several layers of a content representation network; then, given a point q_i in the source image feature space $\boldsymbol{c}_x$, we compute the the spatial correlation map (SCM) as $\boldsymbol{S}_x^i = (\boldsymbol{c}_x^{q_i})^T(\boldsymbol{c}_x^{q_{pi}})$, where $\boldsymbol{c}_x^{q_i} \in \mathbb{R}^{C\times 1}$ is the point feature with C channels, $\boldsymbol{c}_x^{q_{pi}} \in \mathbb{R}^{C\times N_p}$ are the point features within the surrounding patch of N_p points centered at q_i and $\boldsymbol{S}_x^i \in \mathbb{R}^{1\times N_p}$. The content-preserving loss is the cosine distance between N_s SCMs of $\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$:

$$\mathcal{L}_s = \left\|1 - cos(\boldsymbol{S}_x, \boldsymbol{S}_{\hat{y}})\right\|, \quad (3)$$

where $\boldsymbol{S}_x, \boldsymbol{S}_{\hat{y}} \in \mathbb{R}^{N_s\times N_p}$. We use the learned content representation instead of the fixed one, which is trained in a self-supervised contrastive learning fashion to make the extracted content features adapt to the medical image domain.

We train the SR-augmented generator G alternatively with a discriminator D by the following adversarial loss:

$$\mathcal{L}_D = -\mathbb{E}\left[logD(\boldsymbol{y})\right] - \mathbb{E}\left[log\left(1 - D(G(\boldsymbol{x}, \boldsymbol{y}))\right)\right], \quad (4)$$

$$\mathcal{L}_G = \mathbb{E}\left[log\left(1 - D(G(\boldsymbol{x}, \boldsymbol{y}))\right)\right] + \alpha\mathcal{L}_s, \quad (5)$$

where α is a hyperparameter trading off between sample-specific style transferring and content preserving.

3.2. Cascaded ViT Registration Network

We decompose large deformations between the fixed image tMR ($\boldsymbol{f}$) and the moving image cMR ($\boldsymbol{m}$) as global

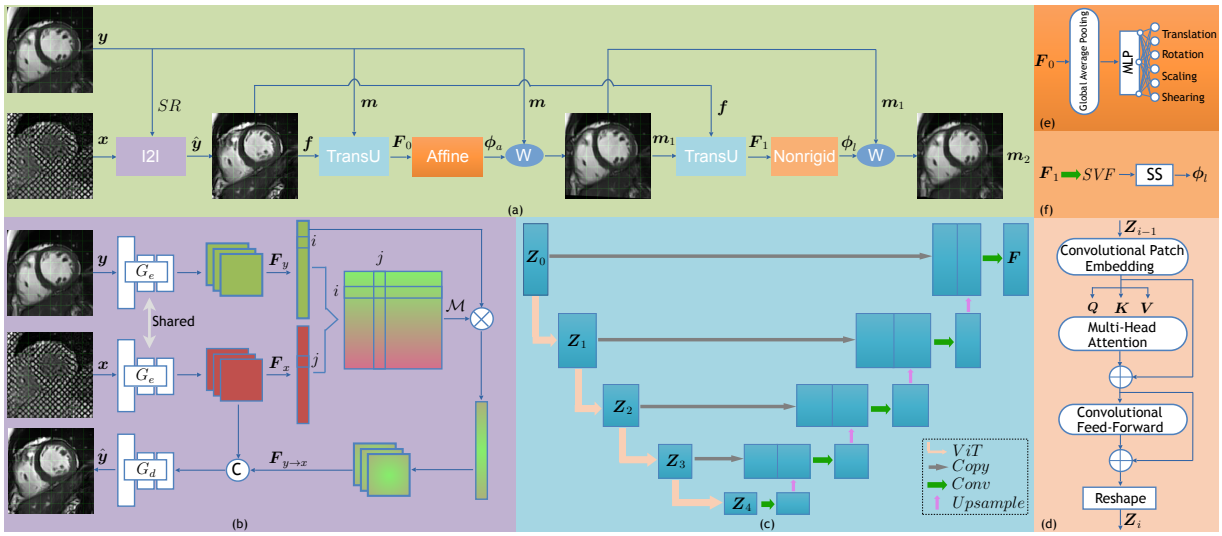

Figure 2. (a) The pipeline of ECaT for multi-modal image registration. (b) The generator of the style reference-augmented I2I network. $\mathcal{M}$ is the cross-domain correspondence correlation matrix. (c) The shared ViT-based feature extractor (TransUnet) by affine and non-rigid registration stages. (d) ViT used in TransU. (e) Affine registration head. (f) Non-rigid registration head. "SR" means style reference; "W" is "warp"; "C" is channel-wise concatenation; "SVF" is a stationary velocity field; "SS" is the scaling and squaring layer.

affine and local non-rigid deformation components, and propose a novel cascaded ViT-based registration network $ViTR$ as shown in Fig. 2 (a). Previous work uses either multi-resolution [29] or cascaded networks [48] to predict large deformations. The cascaded method is more efficient for the use of a single, shared subnetwork, which can reduce network parameters significantly. However, nearly all previous methods use a shrinking network to estimate affine motion while a shrinking-expanding network to estimate non-rigid motion, which makes it impossible to share a common subnetwork for both kinds of deformations. We first introduce a TransUnet (Fig. 2 (c)) as the shared subnetwork for efficient feature embedding. Then, for affine and non-rigid deformation estimation, the TransUnet is coupled with their own heads (Fig. 2 (e) and (f)). Finally, the two subnetworks are cascaded as an end-to-end architecture $ViTR$.

The TransUnet consists of an encoder and a decoder. Compared with pure convolution-based U-Net [36], it differs by the ViT-based encoder. As shown in Fig. 2 (d), we employ the ViT introduced in [30]. It replaces the linear patch embedding with the convolutional patch embedding and adds a depthwise convolution layer in between the two hidden layers of a multilayer perceptron (MLP) block in the feed-forward layer. These two improvements can add more locality into the ViT. Therefore, it efficiently models not only the long-range dependencies within the image patches, by the self-attention mechanism, but also the relationship between a certain patch and its neighbours, by the locality. However, this ViT is a shrinking network dedicated for affine registration tasks. To fit it in the local non-rigid deformation estimation, we adopt the ViT to replace the convolution layers in the encoder of the original U-Net and make it as a local and long-range dependency modeling layer:

$$\boldsymbol{Z}_i = ViT(\boldsymbol{Z}_{i-1}), \tag{6}$$

where $\boldsymbol{Z}_i$ is the i-th layer's feature embedding and $\boldsymbol{Z}_0 = (\boldsymbol{f}, \boldsymbol{m})$. Note, we use $stride = 2$ in each convolutional patch embedding layer to downsample the embedding size. With the ViT embeddings of the input image pair, we use a convolutional decoder to upsample them and further model local dependencies among feature embeddings. The upsampling layers in the decoder make it possible to learn the positional information of patch embeddings implicitly [6], hence we eliminate the positional embedding layer in the original ViT. The skip connection between the encoder and the decoder further enhances the local and long-range dependency modeling of feature embeddings at each scale. Finally, the TransUnet outputs the feature embeddings of input fixed and moving image pair: $\boldsymbol{F}_0 = TransU(\boldsymbol{f}, \boldsymbol{m})$, where $\boldsymbol{F}_0 \in \mathbb{R}^{H \times W \times d}$ and d is the embedding dimension.

The affine head consists of a global average pooling layer and a two-layer MLP. We first average $\boldsymbol{F}_0$ along the H and W dimension and then use the MLP to map it to the affine registration parameters. We decouple the affine transformation into four subtransformations, i.e., translation $\boldsymbol{t}$, rotation $\boldsymbol{r}$, scaling $\boldsymbol{s}$ and shearing $\boldsymbol{h}$: $[\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{s}, \boldsymbol{h}] = affine(\boldsymbol{F}_0)$, where $\boldsymbol{t}, \boldsymbol{s} \in \mathbb{R}^2$, $\boldsymbol{r}, \boldsymbol{h} \in \mathbb{R}$. The affine matix $\mathcal{A}$ is given by $\mathcal{A} = \mathcal{T} \cdot \mathcal{R} \cdot \mathcal{S} \cdot \mathcal{H}$, where $\mathcal{T}$, $\mathcal{R}$, $\mathcal{S}$, $\mathcal{H}$ are the translation, rotation, scaling and shearing transformation matrices derived from the above corresponding transformation param-

eters, respectively. With $\mathcal{A}$, the moving image is warped towards the fixed image which gives $\boldsymbol{m}_1 = \boldsymbol{m} \circ \boldsymbol{\phi}_a$, where $\boldsymbol{\phi}_a$ is the affine deformation field derived from $\mathcal{A}$.

The non-rigid head consists of a convolution layer and a scaling and squaring (SS) layer [2, 9]. We first use the convolution layer to map $\boldsymbol{F}_1 = TransU(\boldsymbol{f}, \boldsymbol{m_1})$ to a stationary velocity field (SVF) and then use the SS layer with 7 recurrences to compute the local deformation field $\boldsymbol{\phi}_l$: $\boldsymbol{\phi}_l = nonrigid(\boldsymbol{F}_1)$, where $\boldsymbol{\phi}_l$ is diffeomorphic with the effect of the SS layer. With $\boldsymbol{\phi}_l$, we warp the moving image $\boldsymbol{m}_1$ towards the fixed image, which gives $\boldsymbol{m}_2 = \boldsymbol{m}_1 \circ \boldsymbol{\phi}_l = \boldsymbol{m} \circ \boldsymbol{\phi}_a \circ \boldsymbol{\phi}_l = \boldsymbol{m} \circ \boldsymbol{\phi}$, where $\boldsymbol{\phi} = \boldsymbol{\phi}_a \circ \boldsymbol{\phi}_l$ is the composed deformation field. By minimizing the dissimilarity between $\boldsymbol{f}$ and $\boldsymbol{m}_2$, we could register them: $\mathcal{L}_{sim} = -sim(\boldsymbol{f}, \boldsymbol{m} \circ \boldsymbol{\phi})$, where $sim(\cdot)$ is a similarity measure. Since the I2I module translates the fixed image to the same modality as the moving image, we could choose mono-modal similarity measures, such as L_1, L_2 and normalized local cross-correlation (NCC). In this work, we choose NCC as the similarity measure for its robust performance of medical image registration. The definition of NCC of an image pair I and J is

$$NCC(I, J) = \sum_{\boldsymbol{p} \in \Omega} \frac{\left(\sum_{\boldsymbol{p}_i \in \mathcal{W}} (I(\boldsymbol{p}_i) - \bar{I}(\boldsymbol{p}))(J(\boldsymbol{p}_i) - \bar{J}(\boldsymbol{p})) \right)^2}{\left(\sum_{\boldsymbol{p}_i \in \mathcal{W}} (I(\boldsymbol{p}_i) - \bar{I}(\boldsymbol{p}))^2 \right) \left(\sum_{\boldsymbol{p}_i \in \mathcal{W}} (J(\boldsymbol{p}_i) - \bar{J}(\boldsymbol{p}))^2 \right)}, \tag{7}$$

where $\bar{I}(\boldsymbol{p})$ and $\bar{J}(\boldsymbol{p})$ are the local mean of I and J at position $\boldsymbol{p}$, respectively, calculated in a w^2 window $\mathcal{W}$ centered at $\boldsymbol{p}$, and $\Omega \subset \mathbb{R}^2$ is the 2D image spatial domain. We set $w = 9$ in our experiments. A higher NCC indicates better registration, so the similarity loss between $\boldsymbol{f}$ and $\boldsymbol{m} \circ \boldsymbol{\phi}$ could be $\mathcal{L}_{sim}(\boldsymbol{f}, \boldsymbol{m} \circ \boldsymbol{\phi}) = -NCC(\boldsymbol{f}, \boldsymbol{m} \circ \boldsymbol{\phi})$. We also add a smoothness regularization loss to make the learned local deformation field smooth which is a penalty on its gradients: $\mathcal{L}_{smooth} = \|\bigtriangledown \boldsymbol{\phi}_l\|$. In sum, we train the affine and non-rigid registration network $ViTR$ with the loss

$$\mathcal{L}_{Reg} = \mathcal{L}_{sim} + \lambda \mathcal{L}_{smooth}, \tag{8}$$

where λ is a balancing hyper-parameter.

3.3. Two-stage Training Scheme

To get the deformation field $\boldsymbol{\phi}$ between the moving image $\boldsymbol{m}$ and the fixed image $\boldsymbol{f}$, we sequentially compose the I2I module and the registration module together; the fixed image $\boldsymbol{f}$ will be the fake image $\hat{\boldsymbol{y}}$. In spite of the content-preserving loss in Eq. (3), if we train generator G and registration network $ViTR$ jointly, as in RegGAN [19], the content distortion problem of G still exists because of the trivial solution [21]: $G(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y}$, $\boldsymbol{\phi} = Id$, i.e., the identity transform. We thus train G and $ViTR$ in a two-stage fashion to decouple the I2I task and the registration task.

4. Experiments

4.1. Dataset and Pre-processing

We collected a clinical cardiac MR dataset which consists of 23 subjects' whole heart scans. Paired cMR and tMR scans for each patient are included in the dataset, but they are generally not aligned with each other, due to imaging condition changes during separate scans. Each scan set covers the long-axis (LAX) and short-axis (SAX) views. For the LAX views, it has the 2-, 3-, 4-chamber views. The SAX views include multiple slices from the base to the apex of the left and right heart ventricles. Each set has approximately 10 2D slices, each of which covers a full cardiac cycle forming a 2D sequence. In total, there are 223 2D paired sequences in our dataset. For each cMR sequence, the frame number is 25, while for tMR sequences, the frame numbers vary from $16 \sim 25$. We first used the scan information in the DICOM header to do the rigid registration (translation and rotation) between each pair of cMR and tMR images and then used temporal nearest sampling to resample the tMR image sequences as a fixed frame number of 25. A region of interest (ROI) was extracted from the images to cover the heart, then we resampled them to the same in-plane spatial size 192×192. Each image pair was used as input to the model to estimate the remaining deformation apart from the rigid one; in total, we have 5,575 cMR and tMR pairs in the dataset. We randomly split the dataset into $3,500$, 750 and $1,325$ pairs as the training, validation and test sets, respectively (without patient crossing). For each 2D image value normalization, we first divided them with 2 times the median intensity value of the image and then truncated all the values to be $[0, 1]$. During training, we also did on-the-fly data augmentation with random translation, scaling, rotation and Gaussian noise addition for each image pair.

4.2. Evaluation Metrics

We evaluated both the quality of the translated fake cMR image and the accuracy of the multi-modal registration. For the former evaluation, we used the normalized mean absolute error (NMAE), peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics (see their definitions in [43]); the ground truth cMR image is the warped result by our registration method, since there is no ground truth cMR image corresponding to the tMR image, due to the deformation between the two scans. For the latter evaluation, we asked two clinical experts to annotate the segmentation mask of the Myo wall on the image pair and double check all the annotations. During evaluation, we input the Myo masks on the cMR images and warped them by the estimated motion field $\boldsymbol{\phi}$. With the warped cMR Myo mask and the tMR Myo mask, we computed the Dice score [4] and the 95th quantile Hausdorff distance (HD) score [16]. We also evaluated the diffeomorphic property of the estimated

motion field ϕ by using the percentage of non-positive Jacobian determinant $det(J_\phi)$ [9] pixels on the image plane.

4.3. Baseline Methods

For the I2I task, we compared our method with CycleGAN [54], UNIT [22], MUNIT [15], NICEGAN [8], and RegGAN [19]. Note that, for RegGAN, we used a non cycle-consistency with registration scheme based on the NICEGAN, since it has the best performance [19]. For the multi-modal registration task, we compared our method with the traditional MIND method based on a 3-level multi-resolution iterative optimization scheme [14], and recently proposed cutting-edge deep learning-based unsupervised medical image registration methods VM [4], VM-dif [9] and MIDIR [35]. VM is a deformable registration model, while VM-dif is a diffeomorphic registration model. MIDIR is also a diffeomorphic registration model, but it further uses B-spline free-form deformation (FFD) [37] to parameterize the SVF. We used the online official implementation code to train VM, VM-dif and MIDIR from scratch, following the optimal hyper-parameters suggested by the authors, with NCC or normalized mutual information (NMI) introduced in [35] as the similarity loss.

4.4. Implementation

We implemented our method with PyTorch. For the I2I module, the architecture is the same as in [50], except that, in the generator, we added the cross-domain correspondence module which is the same as in [45]. For the registration module, the ViT architecture is the same as in [30] and the remaining U-Net architecture is the same as VM. Our code will be available on a public github repository if the paper is accepted. We used the Adam optimizer to train the I2I and registration modules with learning rates of $1e^{-4}$ and $5e^{-4}$, respectively. We set hyper-parameters $\alpha = 10$, $\lambda = 5$, via grid search. All models were trained using an NVIDIA Quadro RTX 8000 GPU.

4.5. Results

4.5.1 Quality of tMR to cMR Translation

Method	NMAE ↓	PSNR ↑	SSIM ↑
CycleGAN	0.064 ± 0.018	22.966 ± 6.709	0.551 ± 0.069
UNIT	0.062 ± 0.014	24.529 ± 9.011	0.539 ± 0.073
MUNIT	0.057 ± 0.013	24.378 ± 7.515	$0.562 \pm \mathbf{0.055}$
NICEGAN	0.050 ± 0.011	25.085 ± 7.314	0.613 ± 0.060
RegGAN	0.051 ± 0.011	25.207 ± 7.076	0.601 ± 0.061
Ours	$\mathbf{0.031 \pm 0.011}$	$\mathbf{27.516 \pm 5.938}$	$\mathbf{0.781} \pm 0.076$

Table 1. Average NMAE, PSNR (dB), SSIM for tMR to cMR image translation.

We show the results of tMR to cMR translation in Table 1. Our method outperforms all the baseline methods for the NMAE, PSNR and SSIM metrics by a large margin. We also show an image translation example in Fig. 3. We can see clearly that all the baseline methods fail to preserve the content of the input tMR image. As shown in the area highlighted by the blue box near to the Myo wall which is our ROI, if the translated image has content distortion, the downstream registration performance will be harmed. While our method can successfully obtain content-preserving I2I results, we also notice that, the translated image has a better style coherence with the cMR image to be registered, than the results of the baseline methods. As indicated by the yellow box, having an SR image input to the generator, our method leverages the style of each specific SR and directly transfers it into the translation result. Note that, we use the image to be registered as the SR image, which might be unaligned with the source image. However, with the cross-domain correspondence module and a content-preserving loss, our method can learn to align the SR image features with the source features, indicating by the good alignment between the correlation matrix warped SR image in Fig. 3 (c) and the source image in Fig. 3 (a).

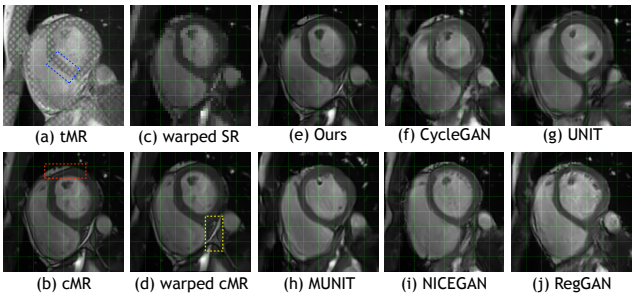

Figure 3. Results of tMR to cMR translation. (a) tMR; (b) cMR (also as the style reference); (c) warped style reference by $\mathcal{M}$; (d) warped cMR by ϕ; (e)∼(j) translated fake cMR from (a). Red box shows unaligned region of cMR with tMR. Blue box shows content distortion region for baseline methods. Yellow box highlights area where baseline methods fail to keep style coherence.

4.5.2 Accuracy of tMR and cMR Registration

We show the results of cMR to tMR registration in Table 2. We also show a registration example in Fig. 4. See more detailed results in the Supplementary Material. After rigid registration using the DICOM header, rigid motion due to scan positioning change could be corrected, which is demonstrated by the close locations of the liver dome indicated by the blue lines. However, the low Dice score after rigid registration suggests that large remaining deformations still exist between the two separate scans.

For all the learning-based baseline methods, the NMI similarity metric-based models outperform the NCC-based ones, suggesting that for multi-modal registration, NMI is

Method	Dice (%) ↑	HD (mm) ↓	$det(J_\phi) \leqslant 0$ (%) ↓	Time (s) ↓
Rigid	63.4 ± 16.3	4.35 ± 2.90	–	–
MIND	72.6 ± 14.7	3.17 ± 2.36	$\mathbf{0.00 \pm 0.00}$	7.428 ± 2.457
VM (C)	63.7 ± 15.5	4.09 ± 2.78	3.80 ± 0.93	$\mathbf{0.003} \pm 0.045$
VM (I)	72.0 ± 13.3	3.46 ± 2.41	0.03 ± 0.05	0.003 ± 0.045
VM-dif (C)	66.0 ± 14.9	4.05 ± 2.81	0.03 ± 0.01	0.005 ± 0.048
VM-dif (I)	72.3 ± 13.0	3.37 ± 2.16	0.12 ± 0.33	0.005 ± 0.040
MIDIR (C)	67.8 ± 15.4	4.03 ± 2.92	0.00 ± 0.00	$0.008 \pm \mathbf{0.040}$
MIDIR (I)	72.7 ± 13.0	3.32 ± 2.13	0.10 ± 0.29	0.009 ± 0.053
Ours	$\mathbf{77.4 \pm 11.9}$	$\mathbf{2.54 \pm 1.89}$	0.01 ± 0.01	0.068 ± 0.174

Table 2. Average Dice, Hausdorff distances (HD), percentage of pixels with non-positive Jacobian determinant on the image plane and running time for multi-modal cMR and tMR image registration. 'C' means 'NCC', 'I' means 'NMI'.

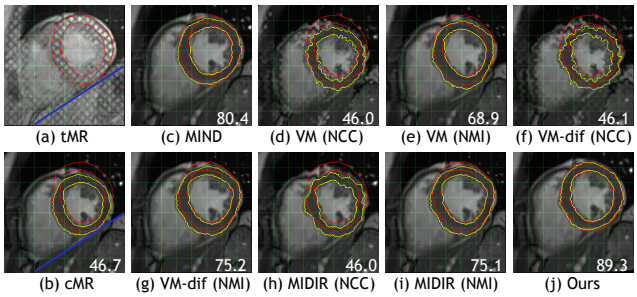

Figure 4. Results of cMR registration to tMR. (a) tMR; (b) cMR after rigid registration; (c)∼(j) warped cMR from (b). Red/yellow contour shows ground truth/warped Myo wall on tMR/warped cMR. The right bottom of each cMR image shows the Dice score.

more suitable than NCC. The MIND method achieves relatively good performance by using the MIND features-based similarity metric and multi-resolution (3 levels) optimization to cope with large deformations. Our method is much more accurate than MIND since the cascaded ViT can predict both the affine and non-rigid deformations. Another observation is that the diffeomorphic registration models perform better than deformable models, due to the smoothness and invertibility property of the diffeomorphic motion fields. The MIDIR model improves the registration performance from VM-dif model by using B-spline FFD parameterization of the SVF, which makes the diffeomorphic motion field smoother. Our registration module is simpler than MIDIR, since we have no B-spline FFD parameterization of the SVF. We benefit from the diffeomorphic registration model and the NCC similarity metric, which maintain the average portion of our model's non-positive Jacobian determinants on the image plane close to zero and ensure the learned deformation field is close to a one-to-one mapping.

4.5.3 Running Time Analysis

In Table 2, we report the average inference time for tMR and cMR image registration by using an Intel Xeon CPU and an NVIDIA Quadro RTX 8000 GPU for different methods. While the unsupervised deep learning-based methods utilize both CPU and GPU during inference, the conventional method (MIND) only uses the CPU. Clearly, the learning-based method is much faster than the conventional iteration-based MIND method. Our method can complete the inference of an image pair registration in far less than one second. Although our method is slower than other learning-based methods, we additionally obtain the translated fake cMR images, which could be used in other tasks, such as cross-domain image segmentation.

4.6. Ablation Study

4.6.1 Efficacy of Content-Preserving Loss and Enforced Style Consistency

Model	$\mathcal{L}_s$	SR	I2I	Registration
			PSNR (dB) ↑	Dice (%) ↑
A1			25.039 ± 7.720	74.0 ± 13.1
A2	✓		24.028 ± 6.004	74.7 ± 13.7
A3		✓	$\mathbf{29.706} \pm 9.931$	72.0 ± 13.6
Ours	✓	✓	$27.516 \pm \mathbf{5.938}$	$\mathbf{77.4 \pm 11.9}$

Table 3. Results of ablation study for content-preserving loss and style reference for unsupervised sample-specific style learning.

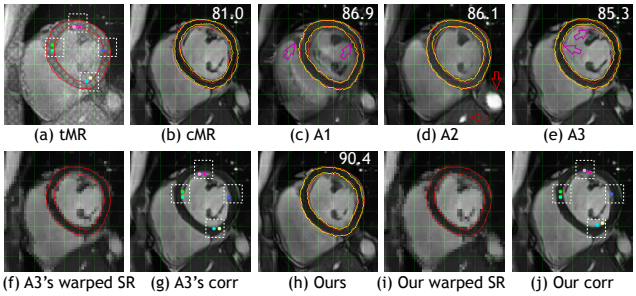

Figure 5. Results of cMR and tMR translation and registration. (a) tMR and query points; (b) cMR; (c)∼(e), (h) fake cMR translated from (a); (f), (i) warped style reference by $\mathcal{M}$; (g), (j) learned correspondences in cMR for the query points in (a). Red/yellow contour shows ground truth/warped Myo wall on tMR/(fake) cMR. The right top of each (fake) cMR image shows the Dice score.

To verify the efficacy of components in our I2I module, we trained different models shown in Table 3. With these I2I modules, we then trained their corresponding registration modules ($ViTRs$) in the second stage to test the influence of the I2I module on downstream registration. We show an visual example in Fig. 5. See more detailed results in the Supplementary Material. From Table 3 and Fig. 5, without the content-preserving loss (A1 and A3), content distortion of the translated fake image is unavoidable (pink arrows). A3 has the SR input into the generator

and gives the best I2I quality under the three metrics. However, the output fake image has the most severe content distortion, thus deteriorating the registration performance the most. As shown in the white boxes and the warped style references in Fig. 5, while A3 fails to learn plausible cross-domain correspondences, our model can learn correct cross-domain correspondences without supervision and generate both content-preserving and style-coherent fake cMR images. A2 is the same as the LSeSim model in [50], which is content-preserving, but it cannot learn the proper style of each cMR image to be registered without the SR (red arrows). By comparing our model with A2, we note how the enforced sample-specific style consistency during I2I can significantly boost the downstream registration accuracy.

4.6.2 Efficacy of ViT for Registration

To verify the efficacy of the ViT-based embedding for registration, we replaced it with the pure convolutional layer turning the TransUnet into a U-Net and named model B1. We trained B1 with the same generator as our model which was trained in the first stage. From Table 4, with the long-range dependency modeling ability of ViT, it outperforms B1. Although B1 performs better for the affine registration task than ours, the non-rigid registration can benefit more from the ViT than the CNN. Besides, the design of cascaded ViTs in our method ensures a global optimum which predicts a better composed deformation field. Also note that, the baseline registration models in Table 2 only have the non-rigid registration part. Our style reference-augmented I2I module could learn sample-specific style of each cMR image to be registered, making the downstream mono-modal registration more accurate. Even only with the non-rigid registration, both B1 and our model significantly outperform the baseline methods in Table 2.

Model	CNN	ViT	Dice (%) ↑		
			Affine	Non-rigid	Composed
B1	✓		**70.5 ± 14.4**	75.0 ± 13.7	76.7 ± 12.5
Ours		✓	69.6 ± 14.5	**75.4 ± 13.5**	**77.4 ± 11.9**

Table 4. Ablation study results of CNN- and ViT-based feature embedding for the registration module.

4.6.3 Efficacy of Shared Feature Embedding for Cascaded Affine and Non-rigid Registration

To study the efficacy of shared TransUnet for feature embedding during cascaded affine and non-rigid registration, we first created models C1 and C2, in which the TransUnet was replaced by only an encoder of TransUnet (TransEncoder) for affine registration. Note that the TransEncoder was unshared with the encoder of subsequent TransUnet for non-rigid registration in C1, but shared in C2. Then we created model C3, in which two different TransUnets were used for the two stages of registration. We trained C1, C2, C3 with the same generator as our model which was trained in the first stage. From Table 5, we can note that TransUnet is more efficient than TransEncoder for cascaded affine and non-rigid registration. We also note that the shared embedding models can not only significantly reduce the learning parameters but also effectively avoid overfitting caused by excess of network parameters.

Model	Affine	Non-rigid	Shared	Params (M) ↓	Dice (%) ↑
C1	TE	TU		0.173	76.7 ± 12.2
C2	TE	TU	✓	0.120	76.9 ± 12.9
C3	TU	TU		0.237	77.0 ± 13.1
Ours	TU	TU	✓	**0.119**	**77.4 ± 11.9**

Table 5. Ablation study results of different feature embedding fashions for cascaded affine and non-rigid registration. 'TE' means 'TransEncoder'; 'TU' means 'TransUnet'.

4.6.4 Efficacy of Two-stage Training

From Table 6, with the content-preserving loss and the style reference input, if we train the generator and the registration network jointly (D1), content distortion could occur in the translated fake image and impair the downstream registration performance. Two-stage training decouples the I2I task and registration task, avoiding the trivial solution of the joint training scheme. Our method thus ensures both a robust sample-specific style consistency and accurate multi-modal registration results.

Model	Joint	Two-stage	I2I	Registration
			PSNR (dB) ↑	Dice (%) ↑
D1	✓		27.435 ± **5.683**	76.1 ± 13.1
Ours		✓	**27.516** ± 5.938	**77.4 ± 11.9**

Table 6. Ablation study results of joint and two-stage training schemes for I2I and registration.

5. Conclusion

In this work, we proposed a novel multi-modal medical image registration method. We proposed an unsupervised exemplar-based image-to-image translation module to augment the sample-specific style coherence of the translated fake image with each image to be registered. Further, we proposed a cascaded ViT to estimate large affine and non-rigid deformations between modalities. Extensive experiments on a real clinical tMR and cMR dataset verified the efficacy and efficiency of our method.

Acknowledgments. This research has been partially funded by research grants to D. Metaxas through NSF: 2310966, 2235405, 2212301, 2003874, 1951890 and NIH 2R01HL127661.

References

[1] Mihaela Silvia Amzulescu, M De Craene, H Langet, Agnes Pasquet, David Vancraeynest, Anne-Catherine Pouleur, Jean-Louis Vanoverschelde, and BL Gerber. Myocardial strain imaging: review of general principles, validation, and sources of discrepancies. *European Heart Journal-Cardiovascular Imaging*, 20(6):605–619, 2019. 1

[2] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006. 5

[3] Leon Axel and Lawrence Dougherty. Mr imaging of motion with spatial modulation of magnetization. *Radiology*, 171(3):841–845, 1989. 1

[4] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 5, 6

[5] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 1

[6] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *arXiv preprint arXiv:2111.10480*, 2021. 3, 4

[7] Junyu Chen, Yufan He, Eric C Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*, 2021. 3

[8] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8168–8177, 2020. 6

[9] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018. 5, 6

[10] Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019. 3

[11] Neel Dey, Jo Schlemper, Seyed Sadegh Mohseni Salehi, Bo Zhou, Guido Gerig, and Michal Sofka. Contrareg: Contrastive learning of multi-modality unsupervised deformable image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–77. Springer, 2022. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[14] Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis*, 16(7):1423–1435, 2012. 2, 6

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2, 6

[16] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993. 5

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2

[19] Lingke Kong, Chenyu Lian, Detian Huang, Yanle Hu, Qichao Zhou, et al. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems*, 34, 2021. 5, 6

[20] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[21] Fengze Liu, Jinzheng Cai, Yuankai Huo, Chi-Tung Cheng, Ashwin Raju, Dakai Jin, Jing Xiao, Alan Yuille, Le Lu, ChienHung Liao, et al. Jssr: A joint synthesis, segmentation, and registration system for 3d multi-modal image alignment of large-scale pathological ct scans. In *European Conference on Computer Vision*, pages 257–274. Springer, 2020. 2, 5

[22] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. 2, 6

[23] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. Dynast: Dynamic sparse transformer for exemplar-guided image generation. *arXiv preprint arXiv:2207.06124*, 2022. 2

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[25] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145*, 2018. 2

[26] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997. 2

[27] 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)*, pages 768–783, 2018. 2

[28] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 1

[29] Tony CW Mok and Albert Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–221. Springer, 2020. 4

[30] Tony CW Mok and Albert Chung. Affine medical image registration with coarse-to-fine vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20835–20844, 2022. 3, 4, 6

[31] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 2

[32] Steffen E Petersen, Paul M Matthews, Fabian Bamberg, David A Bluemke, Jane M Francis, Matthias G Friedrich, Paul Leeson, Eike Nagel, Sven Plein, Frank E Rademakers, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance*, 15(1):1–10, 2013. 1

[33] Zhen Qian, Dimitris N Metaxas, and Leon Axel. A learning framework for the automatic and accurate segmentation of cardiac tagged mri images. In *International Workshop on Computer Vision for Biomedical Image Applications*, pages 93–102. Springer, 2005. 1

[34] Chen Qin, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen. Unsupervised deformable registration for multi-modal images via disentangled representations. In *International Conference on Information Processing in Medical Imaging*, pages 249–261. Springer, 2019. 2

[35] Huaqi Qiu, Chen Qin, Andreas Schuh, Kerstin Hammernik, and Daniel Rueckert. Learning diffeomorphic and modality-invariant registration using b-splines. In *Medical Imaging with Deep Learning*, 2021. 6

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

[37] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18(8):712–721, 1999. 6

[38] Zhengyang Shen, Xu Han, Zhenlin Xu, and Marc Niethammer. Networks for joint affine and non-parametric image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4224–4233, 2019. 3

[39] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 2

[40] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997. 2

[41] Junshen Xu, Eric Z Chen, Xiao Chen, Terrence Chen, and Shanhui Sun. Multi-scale neural odes for 3d medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 213–223. Springer, 2021. 2

[42] Zhe Xu, Jie Luo, Jiangpeng Yan, Ritvik Pulya, Xiu Li, William Wells, and Jayender Jagadeesan. Adversarial uni- and multi-modal stream networks for multimodal image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 222–232. Springer, 2020. 2

[43] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I Chang, Yan Xu, et al. Mri cross-modality image-to-image translation. *Scientific reports*, 10(1):1–18, 2020. 5

[44] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, and Changgong Zhang. Marginal contrastive correspondence for guided image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10663–10672, 2022. 2

[45] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8052–8061, 2019. 2, 3, 6

[46] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2, 3

[47] Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning dual transformer network for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–138. Springer, 2021. 3

[48] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10600–10610, 2019. 4

[49] Shengyu Zhao, Tingfung Lau, Ji Luo, I Eric, Chao Chang, and Yan Xu. Unsupervised 3d end-to-end medical image registration with volume tweening network. *IEEE journal of*

biomedical and health informatics, 24(5):1394–1404, 2019. 3

[50] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16407–16417, 2021. 2, 3, 6, 8

[51] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Jianming Zhang, Ning Xu, and Jiebo Luo. Semantic layout manipulation with high-resolution sparse attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[52] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 3

[53] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021. 2

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 6

[55] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017. 2