Unsupervised Multi-Modal Medical Image Registration via Discriminator-Free Image-to-Image Translation

Zekang Chen[1] , **Jia Wei**[1*] and **Rui Li**[2]
[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[2]Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA
csjwei@scut.edu.cn

Abstract

In clinical practice, well-aligned multi-modal images, such as Magnetic Resonance (MR) and Computed Tomography (CT), together can provide complementary information for image-guided therapies. Multi-modal image registration is essential for the accurate alignment of these multi-modal images. However, it remains a very challenging task due to complicated and unknown spatial correspondence between different modalities. In this paper, we propose a novel translation-based unsupervised deformable image registration approach to convert the multi-modal registration problem to a mono-modal one. Specifically, our approach incorporates a discriminator-free translation network to facilitate the training of the registration network and a patchwise contrastive loss to encourage the translation network to preserve object shapes. Furthermore, we propose to replace an adversarial loss, that is widely used in previous multi-modal image registration methods, with a pixel loss in order to integrate the output of translation into the target modality. This leads to an unsupervised method requiring no ground-truth deformation or pairs of aligned images for training. We evaluate four variants of our approach on the public Learn2Reg 2021 datasets [Hering *et al.*, 2021]. The experimental results demonstrate that the proposed architecture achieves state-of-the-art performance. Our code is available at https://github.com/heyblackC/DFMIR.

1 Introduction

Different medical image modalities, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT), show unique tissue features due to the acquisition with different scanners and protocols. Multiple image modalities can be fused to provide combined information, which is known as the process of multi-modal image fusion [Pielawski *et al.*, 2020]. A wide variety of clinical applications rely on the fusion of multi-modal data, e.g., preoperative planning and image-guided radiotherapy [Oh and Kim, 2017]. Since the images need to be aligned first through registration for image fusion, it is of great importance to establish anatomical correspondences among images of different modalities by using multi-modal image registration.

Learning-based registration seeks to predict deformation fields directly from a pair of images by maximizing a predefined similarity metric [Fan *et al.*, 2019]. Supervised or semi-supervised learning strategies use ground-truth deformation fields or segmentation masks in the training phase, and may suffer from the lack of data labeling [Uzunova *et al.*, 2017; Hu *et al.*, 2018]. Since it is extremely time-consuming and laborious to label registration data even for specialists, unsupervised methods have been proposed to overcome this limitation solely by maximizing the image similarity between the target image and the source image. However, the performance of unsupervised methods is highly dependent on the choice of cross-modal similarity metrics. Generally, widespread similarity metrics like the sum of squared differences (SSD) and normalized cross correlation (NCC), which are well-suited for mono-modal registration problems [Balakrishnan *et al.*, 2019; de Vos *et al.*, 2017], perform badly in a multi-modal setting. Typically, unsupervised multi-modal registration approaches use Normalized Mutual Information (NMI) and Modality-Independent Neighbourhood Descriptor (MIND) [Maes *et al.*, 2003; Heinrich *et al.*, 2012]. Since NMI, as a global metric, only measures statistical dependence between two entire images, it is difficult to use it for local image alignment. MIND, on the other hand, is a patch-based image similarity metric, which tends to suffer from severe image deformations and cannot achieve global alignment.

Given the recent success of multi-modal image translation [Huang *et al.*, 2018; Park *et al.*, 2020], an alternative solution for addressing multi-modal registration is to convert the problem to a simpler unimodal task using an image-to-image (I2I) translation framework [Qin *et al.*, 2019]. Specifically, translation-based methods use Generative Adversarial Network (GAN) mode to translate images from source modality to target modality. And the GAN consists of a generator and a discriminator, where the generator learns to generate plausible data and the discriminator penalizes the generator for producing unrealistic results. With the GAN mode, the registration network can be trained with unimodal similarity metrics. However, this GAN-based image translation tends to produce shape inconsistency and

*Corresponding Author

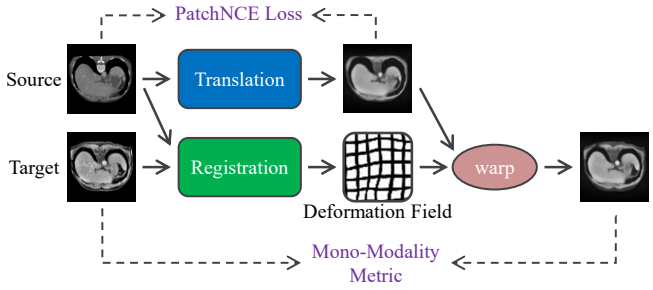


Figure 1: Method overview. Our translation-based registration method learns a cross-modality translation, mapping between the two modalities, which enables the training of the registration network with a mono-modality metric. The source image is warped to align with the target image by the deformation field. We use a PatchNCE loss to encourage the translation network to preserve object shapes and design a pixel loss as the mono-modality metric. The pixel loss not only enables the appearance transfer but also measures the image dissimilarity, which accounts for the training of both translation and registration networks simultaneously.

artificial anatomical features which will, in turn, deteriorate the performance of the registration [Arar *et al.*, 2020; Xu *et al.*, 2020]. More specifically, different modalities have very distinct geometric variances caused by the shape of the imaging bed, the imaging protocols of the scanner, and the field of view. We refer to these variances as **"domain-specific deformations"** [Wang *et al.*, 2021]. We argue that the inconsistency and artifacts are introduced by the discriminator that mistakenly encodes domain-specific deformations as indispensable appearance features and encourages the generator to reproduce the deformations. This tends to create unnecessary difficulty for registration tasks. This paper shows that we can improve the performance of multi-modal image registration by removing the discriminator in image-to-image translation.

In this work, we propose a novel translation-based unsupervised registration approach for aligning multi-modal images. Our main idea is to reduce the inconsistency and artifacts of the translation by removing the discriminator as we discussed above. Specifically, we replace the GAN-based translation network with our discriminator-free translation network (shown in Figure 1). The presented translation network incorporates a patchwise contrastive loss (PatchNCE [Park *et al.*, 2020]) to maintain shape consistency during translation and a pixel loss to integrate the output with the target appearance. Additionally, two novel loss terms are also proposed, one is local alignment loss and the other is global alignment loss. The local alignment loss captures detailed local texture information by modeling the detail-rich image patches. The global alignment loss focuses on the overall shape for the detail-missing generated images. The proposed translation-based registration method coupled with these two losses can achieve local and global alignment and yield more accurate deformation fields.

The main contributions of our work can be summarized as follows:

- We present a discriminator-free I2I translation mode to replace the original GAN-based I2I mode and achieve accurate registration.
- We design a contrastive PatchNCE loss in our translation-based registration model as a shape-preserving constraint.
- We also propose a local and a global alignment losses to further improve the registration performance.

2 Related Work

In recent years, many translation-based multi-modal registration approaches have been proposed. They commonly follow a registration-by-translation framework: an I2I translation network is first trained to synthesize fake images with the target appearance, and then mono-modality metrics can be used in the target domain. In unsupervised translation, cycle-consistent generative adversarial networks (CycleGAN [Zhu *et al.*, 2017]) are widely adopted. However, cycle consistency leads to multiple solutions, which means that the translated images can not maintain the anatomical structure of source images and may contain artifacts [Kong *et al.*, 2021]. With inaccurate image translation, the performance of multi-modal registration tends to be degraded. To address this difficulty, other approaches beyond CycleGAN have been proposed.

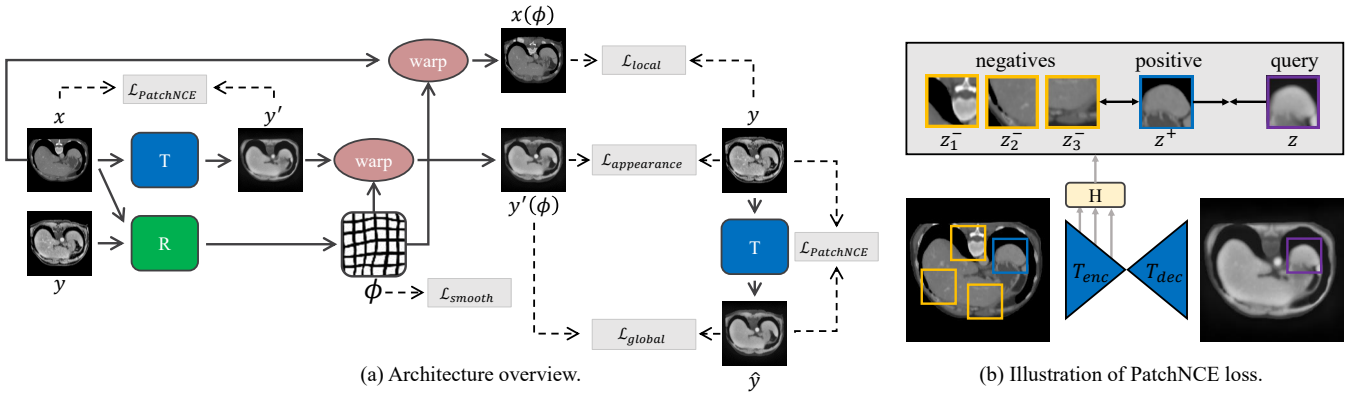
Qin et al. [2019] use image disentanglement to decompose images into common domain-invariant latent shape features and domain-specific appearance features. Then the latent shape features of both modalities are used to train a registration network. But this method still relies on cycle consistency and a GAN mode, which inevitably introduces inconsistency and hampers the process of registration.

Arar et al. [2020] attempt to force the translation and the registration steps to be commutative, which can implicitly encourage the translation network to be structure-preserving. The structure-preserving translation network allows the use of simple mono-modality metrics for training a registration network. However, their method is still GAN-based, which means the structure consistency will be affected by the presence of the discriminator.

Closest to our work, Casamitjana et al. [2021] propose a synthesis-by-registration method, which is different from the previous registration-by-synthesis methods. Their approach is made up of two stages: stage one is training a registration network on pairs of images from the target domain with data augmentation, and stage two is training an I2I network while freezing the parameters of the registration network. However, their registration network is first trained on the images from the target modality instead of images from the two modalities, which may guide the registration network to generate an unrealistic deformation field. And both the registration network and translation network are used in test time in their scheme. On the contrary, our method is a joint framework with end-to-end optimization and only the registration network is needed in test time, which leads to a more reliable deformation field.

3 Method

In this work, we propose an end-to-end learning framework for registering multi-modal image pairs in a fully unsupervised manner. Our core idea is to replace GAN-based I2I translation with our novel discriminator-free I2I translation, shown in Figure 1. The PatchNCE loss maintains shape



(a) Architecture overview. (b) Illustration of PatchNCE loss.

Figure 2: An overview of the proposed method. (a) Architecture overview. Our model consists of two components: a registration network R and a discriminator-free translation network T. The two networks are jointly trained in an end-to-end manner. In our context, the pixel loss $\mathcal{L}_{\text{appearance}}$ is the mono-modality metric computed in the target modality. Based on the architecture presented in Figure 1, we add two novel loss terms $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ to achieve local and global alignments between $x(\phi)$ and y. (b) Illustration of the PatchNCE loss.

consistency and the pixel loss enables the appearance transfer. With the proposed discriminator-free I2I translation, the multi-modal registration task is converted into a unimodal one.

Our method consists of two components: (1) a deformable registration network, and (2) a discriminator-free translation network for mapping images from source domain to target domain and reconstructing images from the target domain. The two components are trained jointly, and only the registration network is used in test time. The pipeline of our method is depicted in Figure 2(a).

We denote the registration network as R and the translation network as T, as in Figure 2. Let $\mathcal{X}$ and $\mathcal{Y}$ denote two paired image domains, where $\mathcal{X}$ is source domain and $\mathcal{Y}$ is target domain. Pairing means that each image $x \in \mathcal{X}$ has a corresponding unaligned image $y \in \mathcal{Y}$ representing the same anatomical structure. The process of registration is to find a deformation field that aligns the source image x to the target image y accurately. Given an image pair (x, y) as input, R learns to predict a deformation field ϕ, which describes how to non-rigidly align x to y. Meanwhile, T takes x and y as the inputs and outputs target-modality images y' and $\hat{y}$, where $y' = T(x)$ and $\hat{y} = T(y)$. y' is the translated image, which has similar appearance as images in domain $\mathcal{Y}$. $\hat{y}$ is the reconstructed image of y. The PatchNCE loss $\mathcal{L}_{\text{PatchNCE}}$ is employed to force y' and $\hat{y}$ to keep the orginial structure of x and y, respectively. $x(\phi)$ and $y'(\phi)$ are warped images from the source image x and the translated image y'. The pixel loss $\mathcal{L}_{\text{appearance}}$ enables the appearance transfer of network T between $y'(\phi)$ and y. In addition, local alignment loss $\mathcal{L}_{\text{local}}$ and global alignment loss $\mathcal{L}_{\text{global}}$ are employed to further improve the registration performance.

3.1 Registration Network

The registration network R takes an image pair (x, y) as an input and outputs a deformation field $\phi = R(x, y)$. The warped image $x(\phi)$ is aligned with y. In a two-dimensional setting, the deformation field is a matrix of 2D vectors, indicating the moving direction for every pixel in the source image x. To generate smooth deformation fields and penalize the tendency of overly distorting the deformed image $x(\phi)$, we adopt an L_2-norm of the gradients of the deformation field as the regularization term [Hoopes *et al.*, 2021], which is denoted as $\mathcal{L}_{smooth}$. Formally, the loss at pixel $v = (i, j)$ is given by:

$$\mathcal{L}_{smooth}(\phi, v) = \sum_{u \in N(v)} \|\phi(u) - \phi(v)\|_2 , \qquad (1)$$

where $N(v)$ denotes a set of neighbor pixels of the v.

3.2 Discriminator-Free Translation Network

Our translation network T takes images from the source domain $\mathcal{X}$ and outputs translated images that have similar appearance as images in the target domain $\mathcal{Y}$. We divide our translation network T into two components, an encoder T_{enc} followed by a decoder T_{dec}, shown in Figure 2(b). T_{enc} extracts shape-related features, while T_{dec} learns to perform shape-preserving modality translation with those features. Given the input x, T_{enc} and T_{dec} jointly generate the output $y' = T(x) = T_{dec}(T_{enc}(x))$.

A key task of our method is to train the I2I translation network T to translate images without the discriminator. If the output from T is shape-preserving, which implies T cannot deform the anatomical structure, the alignment task will be done solely by registration network R. Therefore, we propose the contrastive PatchNCE loss to enforce the shape consistency and the pixel loss to enable the appearance transfer from the source modality to the target modality.

PatchNCE Loss. The PatchNCE loss maximizes the mutual information between input image patches and output image patches, which is based on a noise contrastive estimation framework [Oord *et al.*, 2018]. It makes every output patch similar to the corresponding input patch, while different from the other patches within the input. We use a "query" to refer to an output patch and "positive" for the corresponding input patch and "negatives" for the noncorresponding input patches, shown in Figure 2(b). For similarity

measurement, the encoder part T_{enc} of our translation network with additional two-layer multi-layer perceptron (MLP) is used to map image patches to embedded vectors. Specifically, the query, positive, and N negatives are embedded to K-dimensional vectors z, $z^+ \in \mathbb{R}^K$ and $z^- \in \mathbb{R}^{N\times K}$, respectively. $z_n^- \in \mathbb{R}^K$ denotes the n-th negative patch. We convert the similarity measurement to an (N+1)-way classification problem, where the similarity between the query and other samples are expressed as logits. The cross-entropy loss for multi-class classification is calculated, representing the probability of the positive being selected over N negatives. The formula is given by:

$$\ell\left(z, z^+, z^-\right) = -\log\left[\frac{\exp\left(z\cdot z^+/\tau\right)}{\exp\left(z\cdot z^+/\tau\right)+\sum_{n=1}^{N}\exp\left(z\cdot z_n^-/\tau\right)}\right], \quad (2)$$

where τ is the temperature parameter set to 0.07 in our experiments.

Given an input image, the encoder part T_{enc} will generate multilayer hidden features, which form a feature stack. A spatial location in a layer of the feature stack represents a patch of the input image, with deeper layers corresponding to bigger patches. Let L denote the number of extracted layers from the feature stack. A MLP H_l with two layers is used to map the selected encoder features to embedded representations $\{z_l\}_L = \left\{H_l\left(T_{enc}^l(x)\right)\right\}_L$, where T_{enc}^l represents the features of the l-th selected layer and $l \in \{1, 2, \ldots, L\}$. Let $s \in \{1, \ldots, S_l\}$, where S_l represents the number of spatial locations in each T_{enc}^l. For each spatial location s, the corresponding embedded code is referred to as $z_l^s \in \mathbb{R}^K$, and the features at any other locations are denoted by $z_l^{S\setminus s} \in \mathbb{R}^{(S_l-1)\times K}$. Similarly, the corresponding embedded representations of output image y' are $\{z'_l\}_L = \left\{H_l\left(T_{enc}^l\left(G\left(x\right)\right)\right)\right\}_L$.

With features from multiple layers of the encoder, patchwise noise contrastive estimation can be applied on multiple scales. Multilayer PatchNCE loss is given by:

$$\mathcal{L}_{\text{PatchNCE}}(T, H, X) = \mathbb{E}_x \sum_{l=1}^{L}\sum_{s=1}^{S_l} \ell\left(z'^s_l, z_l^s, z_l^{S\setminus s}\right). \quad (3)$$

In addition, $\mathcal{L}_{\text{PatchNCE}}$ is also used on images from target domain $\mathcal{Y}$, which acts as reconstruction loss. Through the loss $\mathcal{L}_{\text{PatchNCE}}(T, H, Y)$, network T outputs the reconstruction $\hat{y} = T(y)$.

Pixel Loss. Contrastive loss is effective to preserve the shape of the input image x. However, without adversarial loss, we still need to maximize the appearance similarity between a translated image y' and the target domain $\mathcal{Y}$ with a pixel loss. We define the pixel loss with L_1-norm as:

$$\mathcal{L}_{\text{appearance}}(T, R) = \left\|y'(\phi) - y\right\|_1, \quad (4)$$

where $y'(\phi)$ indicates the warped image of y'.

$\mathcal{L}_{\text{appearance}}$ explicitly penalize the absolute intensity differences between $y'(\phi)$ and y'. The combination of $\mathcal{L}_{\text{appearance}}$ and $\mathcal{L}_{\text{PatchNCE}}$ leads to a discriminaotr-free and shape-preserving translation. Note that by minimizing the two losses, the registration network R is trained jointly to predict a deformation field ϕ, which aligns y' to y.

3.3 Local and Global Alignment

To enable R to learn the alignment at the local (patch) level, we propose a variant of the PatchNCE loss. Similarly, we refer to a patch of the warped source image $x(\phi)$ as a "query", while "positive" and "negatives" are corresponding and non-corresponding patch(es) within the target image y, respectively. The warped source image and the target image are mapped to embedded vectors $\{q_l\}_L = \left\{H_l\left(T_{enc}^l(x(\phi))\right)\right\}_L$ and $\{p_l\}_L = \left\{H_l\left(T_{enc}^l(y)\right)\right\}_L$, respectively. The patch-wise noise contrastive estimation is computed between the embedded vectors $\{q_l\}_L$ and $\{p_l\}_L$, which is different from the original PatchNCE. For clarity, we denote this variant loss as $\mathcal{L}_{\text{local}}$:

$$\mathcal{L}_{\text{local}}(R) = \mathbb{E}_{x,y} \sum_{l=1}^{L}\sum_{s=1}^{S_l} \ell\left(q_l^s, p_l^s, p_l^{S\setminus s}\right). \quad (5)$$

Applied on the cross-modality image patches, $\mathcal{L}_{\text{local}}$ encourages the registration network R to learn local alignment as shown in Figure 3. Note that the images produced by our discriminator-free translation network do not contain image texture information. This is beneficial to the extraction of global information, shown in Figure 3. Inspired by this, we further propose a global alignment loss:

$$\mathcal{L}_{\text{global}}(T, R) = \left\|y'(\phi) - \hat{y}\right\|_1. \quad (6)$$

Minimizing $\mathcal{L}_{\text{global}}$ leads to similar style between the generated y' and $\hat{y}$. Meanwhile, the registration network R learns a deformation field ϕ to best align y' to $\hat{y}$.

3.4 Final Objective

Our final objective is as follows:

$$\begin{aligned}\mathcal{L} = \lambda_P \cdot \mathcal{L}_{\text{PatchNCE}}(T, H, X) + \lambda_P \cdot \mathcal{L}_{\text{PatchNCE}}(T, H, Y) \\ +\lambda_A \cdot \mathcal{L}_{\text{appearance}}(T, R) + \lambda_L \cdot \mathcal{L}_{\text{local}}(R) \\ +\lambda_G \cdot \mathcal{L}_{\text{global}}(T, R),\end{aligned} \quad (7)$$

where we set $\lambda_P = 0.25$, $\lambda_A = 1$, $\lambda_L = 0.25$ and $\lambda_G = 1$ in our experiments.

4 Experiments

4.1 Datasets

We evaluated our proposed method on two public datasets. Both of them are obtained from MICCAI Learn2Reg 2021 challenge [Hering *et al.*, 2021], which is a comprehensive registration challenge covering different anatomical structures and modalities. Specifically, the first dataset is for the task of CT-MR thorax-abdomen intra-patient registration, and the other one is for the task of CT lung inspiration-expiration registration.

Thorax-Abdomen CT-MR Dataset. This dataset contains 16 pairs of CT and MR abdomen scans. The annotations on all scans are manual and automatic segmentations of multiple organs. Each scan is a 3D volume in size of $192 \times 160 \times 192$ with 2 mm voxel spacing. After preprocessing the data, e.g., coarse affine registration using the Elastix toolbox [Marstal *et al.*, 2016], we randomly split the dataset into 10/2/4 pairs for

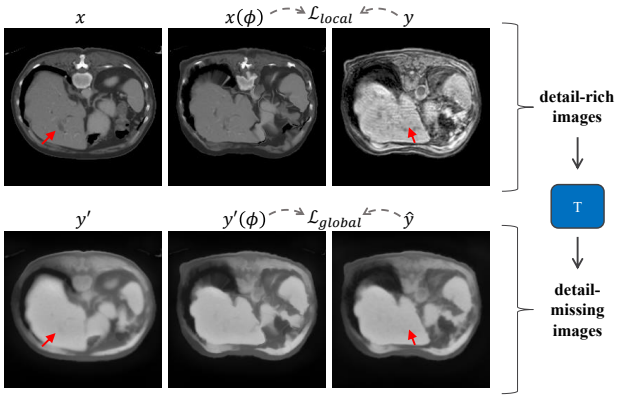


Figure 3: Illustration of local and global alignment losses. Local information such as texture is provided by the detail-rich original image patches (top row). The generated image pair $(y', \hat{y})$ only retains the basic shapes and ignores image texture and details, leading to the focus on global information such as overall shape (bottom row).

train/validation/test, and central 90 slices containing organs in each scan are extracted for our experiments. All slices are padded and resized into 256×256.

Lung CT Inspiration-Expiration (Insp-Exp) Dataset. The dataset consists of 30 pairs of CT lung scans. All lungs are segmented automatically or manually. The challenge in this dataset is to estimate the underlying breathing motion between inspiration and expiration. Each scan is a 3D volume in size of $192 \times 192 \times 208$ with resolution of 1.75 mm. In our experiments, we randomly divide the 30 pairs into 20/4/6 for train/validation/test, and on each volume, we extract middle 100 slices. And each slice is resized into 256×256. To simulate a multi-modal registration task with this single-modal dataset, we synthesize a new modality using the intensity transformation function $\cos(I \cdot \pi/255)$, Gaussian blurring with the kernel size of 3×3 and random elastic deformation sequentially as proposed in [Qin *et al.*, 2019]. Note that the intensity transformation function is only applied to the foreground of each slice. Deformation fields are predicted between real CT inspiration slices and their corresponding synthesized expiration slices.

4.2 Implementation Details

We implement our model based on the framework and implementation of CUT [Park *et al.*, 2020]. The translation network T is a Resnet-based architecture with 9 residual blocks. Our encoder is defined as the first half of the translation network, and five layers of features in the encoder are extracted. The registration network adopts a U-net based architecture with skip connections from contracting path to expanding path [Ronneberger *et al.*, 2015]. For the initialization of networks, we use the Xavier initialization method. Our networks are implemented in PyTorch and all the experiments were conducted on GeForce RTX 2080 Ti. We use Adam optimizer to train our model for 300 epochs with parameters $lr = 0.0002$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Linear learning rate decay is activated after 200 epochs.

	CT → MR		MR → CT	
Method	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
Affine	0.649(0.031)	14.141(2.126)	0.655(0.031)	14.009(2.082)
MIND	0.666(0.039)	14.102(2.446)	0.682(0.024)	14.124(2.320)
CGAN	0.698(0.027)	13.798(2.097)	0.695(0.018)	13.660(1.782)
RGPT	0.713(0.022)	15.844(2.284)	0.717(0.008)	14.706(1.635)
SbR	0.735(0.016)	12.761(1.851)	0.746(0.026)	12.672(2.090)
Ours	**0.772**(0.025)	**12.137**(2.392)	**0.784**(0.025)	**11.889**(2.207)
w/o $\mathcal{L}_{\text{local}}$	0.757(0.027)	12.361(2.609)	0.763(0.022)	12.405(2.045)
w/o $\mathcal{L}_{\text{global}}$	0.763(0.026)	12.453(2.372)	0.770(0.023)	12.232(2.045)
w/o l & g	0.753(0.021)	12.742(2.384)	0.762(0.022)	12.378(1.952)
w/ dis	0.764(0.025)	12.470(2.311)	0.769(0.024)	12.132(2.076)

Table 1: Evaluation of bi-directional multi-modal registration on the Thorax-Abdomen CT-MR dataset in terms of DSC, HD95 (and standard deviation in parentheses).

4.3 Evaluation

Metrics. For the thorax-abdomen dataset, we directly use multi-organ segmentation masks to evaluate the registration accuracy. Dice similarity coefficient (DSC [Dice, 1945]) and 95% percentile of Hausdorff distance (HD95 [Huttenlocher *et al.*, 1993]) are computed between multi-organ masks of a warped source image and its target image. Similarly, DSC and HD95 are calculated between the provided lung masks for the evaluation of the lung CT dataset. DSC is used to measure the accuracy of registration, while HD95 measures reliability. A higher DSC and lower HD95 indicate a better performance of the registration model.

Baselines. We compare our method against several recent state-of-the-art multi-modal registration methods and some other well-established methods. Specifically, the competing methods are: (1) **Affine**, affine registration based on the normalized mutual information using the Elastix toolbox; (2) **MIND**, a VoxelMorph architecture [Balakrishnan *et al.*, 2019] with similarity metric MIND; (3) **CGAN**, a CycleGAN, which is pre-trained on unpaired images, combines with the VoxelMorph registration network using mono-modal similarity metric NCC; (4) **RGPT**, a multi-modal registration model via geometry preserving translation [Arar *et al.*, 2020]; (5) **SbR**, a recent synthesis-by-registration model based on contrastive learning [Casamitjana *et al.*, 2021]. For ablation study, we propose three variants of our model by removing the loss terms one by one: ours w/o $\mathcal{L}_{\text{local}}$, ours w/o $\mathcal{L}_{\text{global}}$, and ours w/o $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ (ours w/o l & g). To investigate if the inaccuracy is introduced by the discriminator, we also study a variant with a discriminator plugged into the translation network in our model (ours w/ dis).

4.4 Results

The quantitative results on the two datasets are summarized in Table 1 and Table 2. We compare the proposed method with the other five methods by measuring the registration accuracy with the DSC and HD95. Our method consistently outperforms other competing methods. Our registration network can predict more accurate deformation fields, even when there exists significant shape deformation and style difference between source images and target images. In addition, Figure 4 displays four examples of the warped source images and their corresponding deformation fields. The deformation

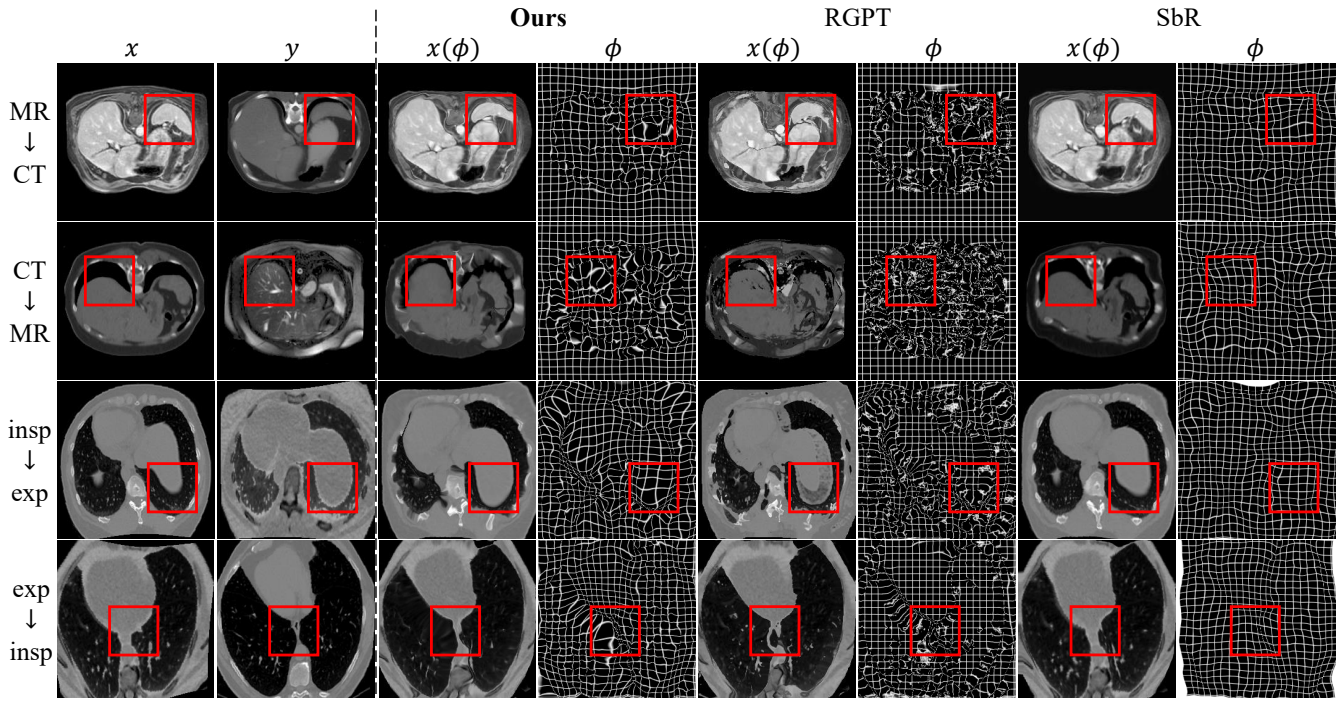


Figure 4: Visualization results of our method against other methods. The unaligned image pair (x, y) is shown in column 1-2 (on the left of the dotted line). x denotes the source image and y denotes the target image. We show the registration results of three methods: Ours, RGPT and SbR. RGPT and SbR are the most recent state-of-the-art methods. Each registration method occupies two columns, the first column showing the warped source image $x(\phi)$ and the second showing the deformation field ϕ.

	insp → exp		exp → insp	
Method	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
Affine	0.850(0.050)	13.985(2.570)	0.848(0.053)	13.746(2.374)
MIND	0.899(0.057)	11.476(1.656)	0.934(0.047)	9.643(2.973)
CGAN	0.879(0.053)	12.989(1.730)	0.913(0.039)	12.038(2.298)
RGPT	0.914(0.034)	12.578(1.158)	0.947(0.030)	11.573(1.351)
SbR	0.924(0.049)	8.866(2.513)	0.946(0.037)	6.739(2.780)
Ours	**0.938**(0.025)	**7.737**(1.791)	**0.966**(0.019)	**4.367**(1.048)
w/o $\mathcal{L}_{\text{local}}$	0.930(0.024)	8.874(1.773)	0.963(0.015)	4.402(1.410)
w/o $\mathcal{L}_{\text{global}}$	0.932(0.029)	8.789(1.936)	0.963(0.019)	5.382(2.409)
w/o l & g	0.910(0.031)	11.714(1.843)	0.957(0.022)	5.642(2.973)
w/ dis	0.929(0.032)	8.372(2.118)	0.959(0.017)	6.232(1.601)

Table 2: Evaluation of bi-directional multi-modal registration on the Lung Insp-Exp dataset in terms of DSC, HD95 (and standard deviation in parentheses).

fields generated by the RGPT method are corrupted by noise and jitter (the red boxes in column 5-6 in Figure 4). This explains its inferior performance in HD95. Even though RGPT can achieve a good boundary alignment, the image quality is degraded after the registration. On the contrary, SbR produces relatively smooth deformation fields but fails to deal with large-scale deformations (the red boxes in column 7-8 in Figure 4). This is because the registration network of SbR is pre-trained on the images from the target modality with randomly generated deformation fields and its parameters are frozen once the pre-training is done. Different from the above methods, ours achieves the most accurate boundary alignment by capturing both local and global information without any discriminator. (the red boxes in column 3-4 in Figure 4).

4.5 Ablation Study

The primary objective of our work is to achieve accurate multi-modal registration. Therefore, we design three variants of the proposed model to investigate the impact of $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$. The results in Table 1 and 2 show that, $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ are complementary, since the proposed method outperform the three ablated versions. What's more, with only one of the two losses, the registration performance still can be improved. To investigate whether the discriminator will introduce inconsistency and degrade the performance of the registration, we design a variant, ours w/ dis. As can be seen in the last row of Table 1 and 2, the registration performance becomes worse than the proposed method.

4.6 Conclusion

We propose a novel discriminator-free and shape-preserving translation network for multi-modal registration, taking advantage of contrastive learning. The registration network successfully gets rid of inconsistency and artifacts introduced by the discriminator. The contrastive loss ensures shape consistency while the pixel loss enables the appearance transfer. Furthermore, we leverage a local and a global alignment losses to achieve local and global alignment, improving the registration accuracy. Finally, we evaluate the proposed method on two open datasets, and show that it outperforms the state-of-the-art methods.

Acknowledgments

This work is supported in part by the Natural Science Foundation of Guangdong Province (2020A1515010717), NSF-1850492 and NSF-2045804.

References

[Arar *et al.*, 2020] Moab Arar, Yiftach Ginger, Dov Danon, Amit H Bermano, and Daniel 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*, pages 13410–13419, 2020.

[Balakrishnan *et al.*, 2019] 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.

[Casamitjana *et al.*, 2021] Adrià Casamitjana, Matteo Mancini, and Juan Eugenio Iglesias. Synth-by-reg (sbr): Contrastive learning for synthesis-based registration of paired images. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 44–54. Springer, 2021.

[de Vos *et al.*, 2017] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 204–212. Springer, 2017.

[Dice, 1945] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[Fan *et al.*, 2019] Jingfan Fan, Xiaohuan Cao, Qian Wang, Pew-Thian Yap, and Dinggang Shen. Adversarial learning for mono- or multi-modal registration. *Medical image analysis*, 58:101545, 2019.

[Heinrich *et al.*, 2012] 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.

[Hering *et al.*, 2021] Alessa Hering, Lasse Hansen, Tony CW Mok, Albert Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *arXiv preprint arXiv:2112.04489*, 2021.

[Hoopes *et al.*, 2021] Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V Dalca. Hypermorph: Amortized hyperparameter learning for image registration. In *International Conference on Information Processing in Medical Imaging*, pages 3–17. Springer, 2021.

[Hu *et al.*, 2018] Yipeng Hu, Marc Modat, Eli Gibson, Wenqi Li, Nooshin Ghavami, Ester Bonmati, Guotai Wang, Steven Bandula, Caroline M Moore, Mark Emberton, et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis*, 49:1–13, 2018.

[Huang *et al.*, 2018] 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.

[Huttenlocher *et al.*, 1993] 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.

[Kong *et al.*, 2021] Lingke Kong, Chenyu Lian, Detian Huang, ZhenJiang Li, Yanle Hu, and Qichao Zhou. Breaking the dilemma of medical image-to-image translation. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[Maes *et al.*, 2003] Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Medical image registration using mutual information. *Proceedings of the IEEE*, 91(10):1699–1722, 2003.

[Marstal *et al.*, 2016] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. Simpleelastix: A user-friendly, multilingual library for medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 134–142, 2016.

[Oh and Kim, 2017] Seungjong Oh and Siyong Kim. Deformable image registration in radiation therapy. *Radiation oncology journal*, 35(2):101, 2017.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Park *et al.*, 2020] 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.

[Pielawski *et al.*, 2020] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Nataša Sladoje. Comir: Contrastive multimodal image representation for registration. *arXiv preprint arXiv:2006.06325*, 2020.

[Qin *et al.*, 2019] 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.

[Ronneberger *et al.*, 2015] 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.

[Uzunova *et al.*, 2017] Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt. Training cnns for image registration from few samples with model-based data augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 223–231. Springer, 2017.

[Wang *et al.*, 2021] Chengjia Wang, Guang Yang, Giorgos Papanastasiou, Sotirios A Tsaftaris, David E Newby, Calum Gray, Gillian Macnaught, and Tom J MacGillivray. Dicyc: Gan-based deformation invariant cross-domain information fusion for medical image synthesis. *Information Fusion*, 67:147–160, 2021.

[Xu *et al.*, 2020] 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.

[Zhu *et al.*, 2017] 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.