Indescribable Multi-modal Spatial Evaluator

Lingke Kong
Manteia Tech
kid_liet@163.com

X. Sharon Qi
University of California, Los Angeles
xqi@mednet.ucla.edu

Qijin Shen
Fuzhou University
qijinshen@foxmail.com

Jiacheng Wang
Xiamen University
jiachengw@stu.xmu.edu

Jingyi Zhang
Xiamen University
zhangjingyi1@stu.xmu.edu.cn

Yanle Hu*
Mayo Clinic Arizona
Hu.Yanle@mayo.edu

Qichao Zhou*
Manteia Tech
zhouqc@manteiatech.com

Abstract

Multi-modal image registration spatially aligns two images with different distributions. One of its major challenges is that images acquired from different imaging machines have different imaging distributions, making it difficult to focus only on the spatial aspect of the images and ignore differences in distributions. In this study, we developed a self-supervised approach, Indescribable Multi-model Spatial Evaluator (IMSE), to address multi-modal image registration. IMSE creates an accurate multi-modal spatial evaluator to measure spatial differences between two images, and then optimizes registration by minimizing the error predicted of the evaluator. To optimize IMSE performance, we also proposed a new style enhancement method called Shuffle Remap which randomizes the image distribution into multiple segments, and then randomly disorders and remaps these segments, so that the distribution of the original image is changed. Shuffle Remap can help IMSE to predict the difference in spatial location from unseen target distributions. Our results show that IMSE outperformed the existing methods for registration using T1-T2 and CT-MRI datasets. IMSE also can be easily integrated into the traditional registration process, and can provide a convenient way to evaluate and visualize registration results. IMSE also has the potential to be used as a new paradigm for image-to-image translation. Our code is available at https://github.com/Kid-Liet/IMSE.

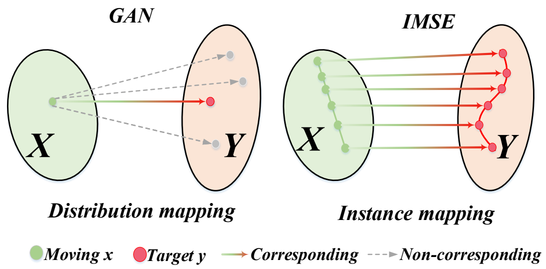

Figure 1. The GAN based methods can only ensure that the distribution of the X domain is mapped that of the Y domain. Ideally, we want to achieve instance registration in which moving and target images are one-to-one corresponded.

1. Introduction

The purpose of multi-modal image registration is to align two images with different distributions (Moving (M) and Target (T) images) by warping the space through the deformation field ϕ. A major challenge in multi-modal image registration is that images from different modalities may differ in multiple aspects given the fact that images are acquired using different imaging machines, or different acquisition parameters. Due to dramatically different reconstruction and acquisition methods, there is no simple one-to-one mapping between different imaging modalities. From the perspective of measuring similarity, mainstream unsupervised multi-modal registration methods can be divided into two categories: similarity operator based registration and image-to-image translation based registration.

Similarity operator based registration uses multi-modal similarity operators as loss functions for registration, for ex-

*Corresponding author.

ample, normalized cross-correlation (NCC) [15, 22, 31, 32], mutual information (MI) [5, 23, 27, 35], and modality-independent neighborhood descriptor (MIND) [3, 8, 13, 39]. Similarity operators are based on a prior mathematical knowledge. These are carefully designed and improved over time. They can be applied to both traditional registration process (Eq. 1) and neural network registration (Eq. 2):

$$\hat{\phi} = \arg\min_{\phi} \mathcal{L}_{sim}\left(M\left(\phi\right), T\right). \tag{1}$$

Or

$$\hat{\theta} = \arg\min_{\theta} \left[\mathbb{E}_{(M,T)}\left[\mathcal{L}_{sim}\left(M, T, g_{\theta}\left(M, T\right)\right)\right]\right]. \tag{2}$$

Similarity operators have several limitations. **1)** It is unlikely to design similarity operators that can maintain high accuracy for all data from various imaging modalities. **2)** It is not possible to estimate the upper limit these operators can achieve and hence it is difficult to find the improvement directions.

Image-to-image translation [1, 16, 17, 21, 29, 38, 42] based multi-modal registration first translations multi-modal images into single-modal images (Eq 3) using a generative adversarial network (GAN [10]), and then use Mean Absolute Error (MAE) or Mean Squared Error (MSE) to evaluate the error at each pixel in space (Eq 4).

$$\begin{aligned}\min_{G}\max_{D} \mathcal{L}_{Adv}\left(G, D\right) = \mathbb{E}_{T}\left[log\left(D\left(T\right)\right)\right] + \\ \mathbb{E}_{M}\left[log\left(1 - D\left(G\left(M\right)\right)\right)\right].\end{aligned} \tag{3}$$

And

$$\hat{\theta} = \arg\min_{\theta} \left[\mathbb{E}_{(M,T)}\left[\|G\left(M\right), T, g_{\theta}\left(G\left(M\right), T\right)\|_{1}\right]\right]. \tag{4}$$

Image-to-image translation based registration cleverly avoids the complex multi-modal problem and reduces the difficulty of registration to a certain extent. However, it has obvious drawbacks. **1)** The methods based on GAN require training a generator using existing multi-modal data. The trained model will not work if it encounters unseen data, which greatly limits its applicable scenarios. **2)** More importantly, registration is an instance problem. However, the method based on GAN is to remap the data distribution between different modal. As shown in Figure 1, the distribution has bias and variance. We cannot guarantee that the translated image corresponds exactly to the instance target image at the pixel level. For example, Figure 2 shows that there is still residual distribution difference between the target image and translated image. Therefore, even if they are well aligned in space, there is still a large error in the same organ.

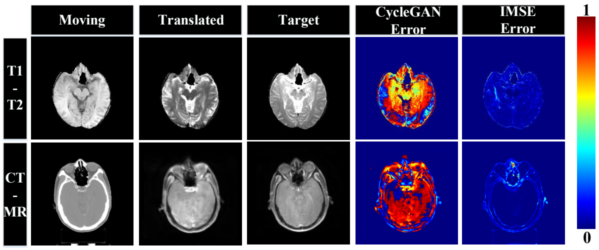

Figure 2. The distributions of the Moving images translated by CycleGAN still have large residual differences from the Target images. IMSE gives smaller error assessment values in the overlapping regions (converging to blue).

To address these challenges, we propose a novel idea based on self-supervision, namely, Indescribable Multi-modal Spatial Evaluator, or IMSE for short. The IMSE approach creates an accurate multi-modal spatial evaluator to metric spatial differences between two images, and then optimizes registration by minimizing the error predicted by the evaluator. In Figure 2, we provide a visual demonstration of IMSE in terms of spatial location for T1-T2 and MRI-CT, respectively. Even though distribution differences between the Moving and Target images are still significant, IMSE is low in overlapping regions of the same organs. The main contributions of this study can be summarized as follows:

- We introduce the concepts of relative single-modal and absolute single-modal as an extension of the current definition of single-modality.

- Based on relative single-modal and absolute single-modal, we propose a self-supervised IMSE method to evaluate spatial differences in multi-modal image registration. The main advantage of IMSE is that it focuses only on differences in spatial location while ignoring differences in multi-modal distribution caused by different image acquisition mechanisms. Our results show that IMSE outperformed the existing metrics for registration using T1-T2 and CT-MRI datasets.

- We propose a new style enhancement method named Shuffle Remap. Shuffle Remap can help IMSE to accurate predict the difference in spatial location from an unseen target distribution. As a enhancement method, Shuffle Remap can be impactful in the field of domain generalization.

- We develop some additional functions for IMSE. **1)** As a measure, IMSE can be integrated into both the registration based on neural network and the traditional registration. **2)** IMSE can also be used to establish a new image-to-image translation paradigm. **3)** IMSE can provide provide a convenient way to evaluate and visualize registration results.

2. Related Work

Multi-modal Similarity Metric: In order to measure the spatial difference between multi-modal images, several classical operators have been proposed. For example, normalized cross-correlation [15, 22, 31, 32] is used to describe the correlation between two vectors or samples of the same dimension, mutual information [5, 23, 27, 35] is used to describe the degree of interdependence between variables, and modal independent neighborhood descriptor [3, 8, 13, 39] is used to describe the local modal characteristics around each voxel. These operators can also be used as loss functions in combination with neural networks [2, 11]. The classical operators are based on mathematical knowledge of researchers. It is difficult to apply to all modal scenes and estimate their upper limits. As GAN [10] becomes popular in image translation task, researchers have proposed various methods based on GAN to translate multi-modal to single-modal images to facilitate registration. Wei et al. [37] used CycleGAN [42] to achieve 2D MR-CT image translation, and converted the 2D slice into 3D volumes through stacking, and finally acted on registration. Qin Chen et al. [30] proposed an unsupervised multi-modal image-to-image synthesis method by separating the latent shape appearance space and content information space. Kong et al. [18] proposed a method to add correction to the image translation to improve the quality of the translated image. These GAN based methods remap the data distribution of different modal. There are always biases and variances within the distribution, so registration is an instance problem.

Domain Generalization: The goal of domain generalization (DG) is to generalize to unseen data by training the model on source domain data [4, 6, 14, 28]. There are methods which aim to learn domain invariant representations by minimizing domain differences between multiple source domains [9, 19, 20, 40]. In addition, several methods handle DG tasks by modifying the normalization layer, such as instance normalization (IN) and batch normalization (BN) [7,26,33,34]. In the medical field, Ziqi Zhou et al. [41] proposed the method of Dual Normalization for segmentation. Wang et al. [36] used a domain knowledge base to store domain specific prior knowledge, and domain attributes to aggregate the characteristics of different domains. Shuffle Remap proposed by us is a pure style enhancement method, which can be easily combined with other domain generalization methods. It is well suited to situations with large deviation scales, such as CT-MR.

3. Methodology

3.1. Motivation

In multi-modal registration, the task is greatly simplified if we can isolate and exclude distribution differences, and focus only on spatial differences. In fact, the GAN-based approach intends to eliminate distribution differences by translating the source-domain image to the target-domain image. However, although the translated image and the target image can be viewed as single-modal relationship, the residual distribution differences may still be significant, making MAE or MSE unsuitable for optimizing registration performance. Further, we can define the relationship between such single-modal images as the **relative single-modal**.

Relative single-modal: For any images $\{x_n\}_{n=1}^{n=N}$, if all satisfy a specific and identical data distribution rule, i.e. $\{x_n\}_{n=1}^{n=N} \sim \mathcal{D}(\mu, \sigma^2)$, then $\{x_1, ..., x_N\}$ are called relative single-modal data with respect to each other. Relative single-modal is a abstract and extensive concept. For example, if both CT and MR belong to the category of medical images compared with natural images, they can be regarded as relative single-modal images; If T1 and T2 belong to the category of MR compared with CT, they can be regarded as relative single-modal images.

Moreover, we can define the relationship between images without residual distribution difference as **absolute single-modal**.

Absolute single-modal: For any $(x_j, x_k) \in \{x_n\}_{n=1}^{n=N}$, where $j \neq k$, if x_j and x_k can be obtained from each other by some particular spatial transformation ϕ only, i.e. $x_j - x_k \circ \phi = 0$ or $x_k - x_j \circ \phi^{-1} = 0$. Then $\{x_1, ..., x_N\}$ is called absolute single-modal data. Absolute single-modal belongs to a narrower concept. To some extent, relative single-modal contains absolute single-modal.

The images that belong to absolute single-modal are completely consistent in modalities. We can regard the absolute-modal difference between two images as a spatial error. Therefore, the question becomes how to make the model measure the spatial error from an absolute single-modal perspective.

3.2. IMSE

The IMSE method involves training of evaluator and registration separately. In this section, we provide a detailed description of the method.

1.Training evaluator: In IMSE, we can completely simulate multi-modal registration data and obtain the absolute single-modal label. We first apply two random spatial transformations T_1 and T_2 to the original image x to obtain the transformed images x_1 and x_2, respectively. The spatial transformation operations include overall rotation, displacement, rescaling and random pixel-wise deformation. Since x_1 and x_2 are from the same image x, they satisfy the definition of absolute single-modal and differ only in spatial location. Once x_1 and x_2 are generated, we subtract the two images to obtain the image of spatial position error, which is used as the label for evaluator training. Next, we add a random noise ε(see Section 3.3 for the specific noise form) to x_1 to create distribution differences between x_1 and x_2.

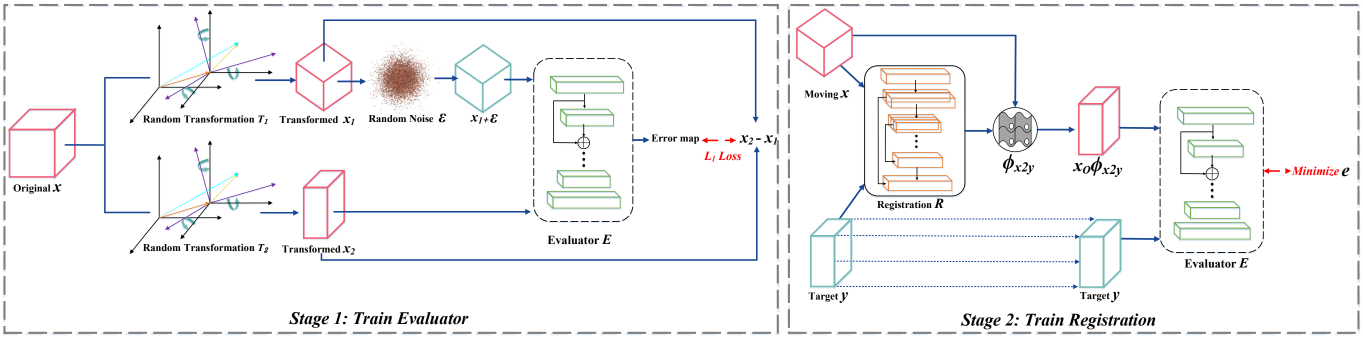

Figure 3. A general overview of the IMSE process. It is divided into two main parts, the training evaluator and the training registration.

$x_1 + \varepsilon$ and x_2 are then stacked according to the number of channels and used as inputs to the evaluator E. Then, the predictions of Evaluator and previously created label are used for training (Eq. 5). It is worth noting that the distribution of $x_1 + \varepsilon$ is arbitrary, but x_2 is unchanged. Therefore, the trained evaluator take the input x_2 as the reference image to predict the absolute single-modal error $(x_2 - x_1)$.

$$\min_{E} \mathcal{L}_{L_1}(E) = \mathbb{E}_{x_1,x_2}\left[\|E(x_2, x_1 + \varepsilon), (x_2 - x_1)\|_1\right]. \tag{5}$$

2.Training registration: Given the moving image x and the target image y, deformation fields ϕ_{x2y} are obtained using a registration network R with x and y as inputs. The deformation fields are used to obtain the warped image $x \circ \phi_{x2y}$ by warping the moving image x. Then, the warped image $x \circ \phi_{x2y}$ and the target images y are fed into the evaluator E to get the spatial error e. By minimizing the error e, we can optimize the parameters of the registration network R (Eq. 6). We should also use a regularization constraint based on the deformation fields (Eq. 7). During the training of the registration network, the parameters of the evaluator remain the same and are not updated.

$$\min_{R} \mathcal{L}_{sim}(R) = \mathbb{E}_{x,y}\left[\|E(x \circ R(x,y), y)\|_1\right]. \tag{6}$$

$$\min_{R} \mathcal{L}_{smooth}(R) = \mathbb{E}_{x,y}\left[\|\nabla R(x,y)\|^2\right]. \tag{7}$$

IMSE uses a neural network to evaluate the similarity between multi-modal images. It is not possible to describe its computational process using mathematical formulas like what traditional similarity operators do. This is analogous to the discriminator in generative adversarial networks. In the training process of the generator, it is impossible to design an analytical operator and use it as a loss function to evaluate the authenticity of the generated images and optimize the generator. To overcome this challenge, the researchers use the classification loss of the discriminator to optimize the generator indirectly. Similarly, IMSE uses the output of the evaluator to update the registration network. Unlike GAN, IMSE has no adversarial process and does not necessarily train both the generator and the discriminator simultaneously as GAN does. IMSE trains the evaluator and registration network separately. Thus, it can save more computational resources.

3.3. Shuffle Remap

The next question we need to consider is what factors determine the performance limitations of IMSE. Based on the training process, IMSE belongs to the category of self-supervision. The label used in the training is absolutely accurate. The upper limit of performance that the evaluator can achieve depends mainly on two factors. First, the degree of deformation added to the image needs to sufficiently cover spatial differences presented in the task. Second, the noise added to the image needs to have sufficient diversity and cover the range of distribution differences.

We propose a simple yet effective style enhancement method named Shuffle Remap to ensure sufficient coverage of distribution differences. Specifically, we first normalize the distribution of the original image X to [-1,1], then generate some random control points in between [-1,1] with the endpoints P_0 and P_N fixed at -1 and 1, respectively. These control points randomly divide the image distribution into N segments. Each segment is assigned an index number n in the order from the smallest to the largest. After that, the segments are randomly disordered and remapped according to the disordered order. For example, Eq. 8 demonstrates the remapping from the range $(Pi, Pi + 1)$ to the range $(Pj, Pj + 1)$ for a given pixel. The complete algorithm flow is given in Algorithm1.

$$x^{'} = \frac{x - p_i}{p_{i+1} - p_i} * (p_{j+1} - p_j) + p_j. \tag{8}$$

Figure 4 shows the effect of Shuffle Remap. We show the results for different segments. Since the index number and location of control points are random, the remapped images either blur the contrast (Remap.1) or enhance the

Algorithm 1 Pseudocode of Shuffle Remap in PyTorch style.

```
# X: the input image and the range is [-1,1]
# r_min: Minimum number of random control points
# r_max: Maximum number of random control points

# number of randomly generated control points
control_point=random.randint(r_min, r_max)
# normalize to the range of the image distribution
dist=torch.rand(control_point)*(1−(−1))+(−1)
# sort from small to large
dist=torch.sort(dist)
# Add endpoint -1 and 1
dist=torch.cat([torch.tensor([−1]), dist])
dist=torch.cat([dist, torch.tensor([1])])
# shuffle the distribution and generate empty new image
shuffle_remap=torch.randperm(control_point+1)
new_X=torch.zeros_like(x)
for i in range(control_point +1):
    target_part=shuffle_remap[i]
    min1, max1=dist[i], dist[i+1]
    min2, max2=dist[target_part], dist[target_part+1]
    # get the coordinates corresponding to the distribution
    coord=torch.where((min1<=X)&(X<max1))
    # Eq.(8)
    new_X[coord]=((X[coord]−min1)/(max1−min1))*
                    (max2−min2)+min2
return   new_X
```

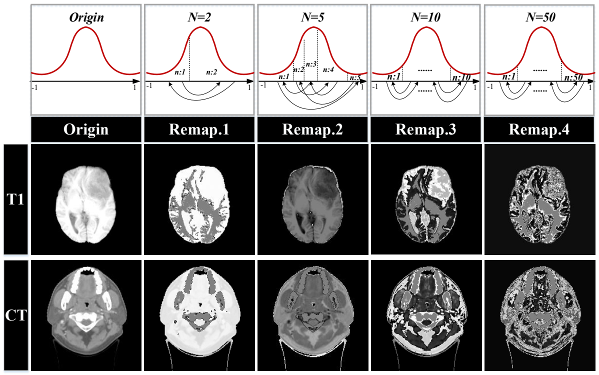

Figure 4. Examples of T1 and CT after Shuffle Remap.

contrast (Remap.3) between the anatomical structures on the original image. The distribution of remapped images also can be very far away (Remap.2) from the original image distribution. In addition, with the increase of segments, the original image can also be confused (Remap.4). Regardless of Shuffle Remap results, the label used in evaluator training is always unique and accurate, which has the obvious benefit of enhancing the model's knowledge of the same anatomical structures while reducing sensitivity to differences in image distribution. Shuffle Remap is very different from the commonly used histogram shift method. Histogram shift only scales the image distribution without changing the relative relationship of the overall image distribution. Shuffle Remap, however, is a completely random remap of the original image distribution.

It is worth noting that we only provide a style enhancement method, and Shuffle Remap is not irreplaceable in the IMSE architecture. Shuffle Remap is a pure style enhancement method, which can be easily combined with other domain generalization methods.

4. Experiments

In this work, we did 4 experiments to evaluate the potential of IMSE. (1) We evaluated the performance of IMSE in multi-modal image registration and compared it with various existing registration methods based on neural networks. (2) We integrated IMSE into the traditional registration procedure. (3) We investigated the feasibility of IMSE as a new image-to-image translation paradigm. (4)We explored the accuracy of using IMSE to assess spatial error.

4.1. Dataset

The first dataset is from T1-T2 modal in BraTS2019 [24], and the second dataset is from clinical CT-MR modal. CT-MR dataset registration accuracy was evaluated based on the parotid gland which was contoured by physicians. Table 1 provides a brief description of the datasets used in this study.

			2D				3D			
Source	**Modality**	**Position**	**Size**	**Train**	**Test**	**Resize**	**Size**	**Train**	**Test**	**Resize**
BraTs 2019	T1-T2	Brain	240×240	18911	640	✗	$48 \times 128 \times 128$	187	20	✓
Clinical	CT-MR	Head neck	192×192	3839	863	✓	$48 \times 128 \times 128$	80	18	✓

Table 1. A brief description of the datasets used in the study.

4.2. IMSE for Registration Based on Neural Network

In this subsection, we compared various registration methods based on neural networks. Baseline uses traditional similarity operators as loss functions to update the registration network, including **NCC** [22], **MI** [27], and **MIND** [13]. There are also GAN based methods which do translation first and then use MAE as the loss function of the registered network, including **CycleGAN** [42], **RegGAN** [17]. We also compared histogram shift using the Bézier curve [25] (**IMSE (BC)**) and Shuffle Remap (**IMSE (SR)**) for style enhancement. The random range of N in Shuffle Remap is [2, 50].

For a fair comparison, all methods used a unified registration network model–VoxelMorph [2]. Please note that in the

Moving→ Target	Methods	2D			3D		
		Dice ↑	HD95 ↓	$\|\nabla\phi\|_2$ ↓	Dice ↑	HD95 ↓	$\|\nabla\phi\|_2$ ↓
	Initial	0.68 ± 0.08	4.17± 1.76	✗	0.81± 0.05	4.13± 1.53	✗
T1 → T2	NCC	0.75 ± 0.05	2.71± 1.33	0.0026	0.84 ± 0.03	3.38 ± 1.03	0.0036
	MI	0.82 ± 0.04	1.70± 1.29	0.0102	0.86± 0.04	2.79 ± 1.44	0.0162
	MIND	0.83 ± 0.04	1.66± 1.27	**0.0023**	0.88± 0.02	2.70 ± 1.01	0.0034
	CycleGAN	0.85 ± 0.03	1.37± 0.95	0.0061	0.88± 0.02	2.94 ± 0.93	0.010
	RegGAN	0.86 ± 0.03	1.25 ± 0.90	0.0091	0.89± 0.01	2.85 ± 0.83	0.0090
	IMSE(BC)	0.83 ± 0.04	1.72 ± 1.30	0.0125	0.84± 0.04	3.21 ± 1.33	0.0105
	IMSE(SR)	**0.89** ± **0.02**	**1.06**± **0.87**	**0.0023**	**0.91**± **0.01**	**2.36** ± **0.77**	**0.0032**
T2 → T1	NCC	0.74 ± 0.04	2.76± 1.35	0.0026	0.84± 0.03	3.07 ± 1.04	0.0038
	MI	0.79 ± 0.05	1.58± 1.31	0.0103	0.88± 0.04	2.33 ± 1.41	0.0166
	MIND	0.81 ± 0.03	2.15± 1.28	0.0023	0.88± 0.02	2.42 ± 1.05	0.0035
	CycleGAN	0.86 ± 0.04	1.19± 0.94	0.0056	0.88± 0.02	2.95 ± 0.95	0.009
	RegGAN	0.86 ± 0.03	1.20± 0.90	0.0071	0.89± 0.01	2.71 ± 0.80	0.0085
	IMSE(BC)	0.80 ± 0.04	1.87 ± 1.34	0.0122	0.85± 0.03	3.01 ± 1.26	0.0097
	IMSE(SR)	**0.89** ± **0.02**	**0.85**± **0.86**	**0.0022**	**0.91**± **0.01**	**2.21** ± **0.75**	**0.0025**

Moving→ Target	Methods	2D			3D		
		Dice ↑	HD95 ↓	$\|\nabla\phi\|_2$ ↓	Dice ↑	HD95 ↓	$\|\nabla\phi\|_2$ ↓
	Initial	0.40 ± 0.07	10.05± 1.93	✗	0.48± 0.05	5.64± 1.41	✗
CT → MR	NCC	0.49 ± 0.04	9.05± 1.75	0.0074	0.54 ± 0.03	5.11 ± 1.18	0.017
	MI	0.50 ± 0.05	8.90± 1.77	0.0075	0.55± 0.03	5.14 ± 1.16	0.019
	MIND	0.50 ± 0.04	8.55± 1.79	0.0019	0.54± 0.02	5.10 ± 1.07	0.008
	CycleGAN	0.56 ± 0.04	8.01± 1.69	0.0022	0.58± 0.02	4.62 ± 0.95	0.010
	RegGAN	0.57 ± 0.03	7.83 ± 1.63	0.0020	0.60± 0.01	4.45 ± 0.90	0.009
	IMSE(BC)	0.50 ± 0.04	8.63 ± 1.69	0.0072	0.55± 0.03	5.08 ± 1.06	0.019
	IMSE(SR)	**0.61** ± **0.02**	**6.92**± **1.51**	**0.0017**	**0.62**± **0.01**	**4.23** ± **0.82**	**0.007**
MR → CT	NCC	0.49 ± 0.05	9.04± 1.75	0.0070	0.57± 0.03	5.01 ± 1.09	0.013
	MI	0.50 ± 0.04	9.18± 1.75	0.0076	0.58± 0.02	5.13 ± 1.15	0.019
	MIND	0.51 ± 0.04	8.86± 1.78	0.0019	0.56± 0.03	5.05 ± 1.19	**0.007**
	CycleGAN	0.58 ± 0.04	7.56± 1.66	**0.0018**	0.58± 0.01	4.81 ± 0.89	0.011
	RegGAN	0.59 ± 0.02	7.30±1.61	0.0021	0.61± 0.01	4.62 ± 0.89	0.009
	IMSE(BC)	0.50 ± 0.04	8.85 ± 1.71	0.0069	0.56± 0.04	4.97 ± 1.16	0.016
	IMSE(SR)	**0.60** ± **0.01**	**7.26**± **1.58**	0.0020	**0.62**± **0.01**	**4.45** ± **0.86**	0.008

Table 2. Registration results of various methods based on the T1-T2 and CT-MR dataset. Initial indicates the results before registration. The source data used to train the IMSE were T1 and CT.

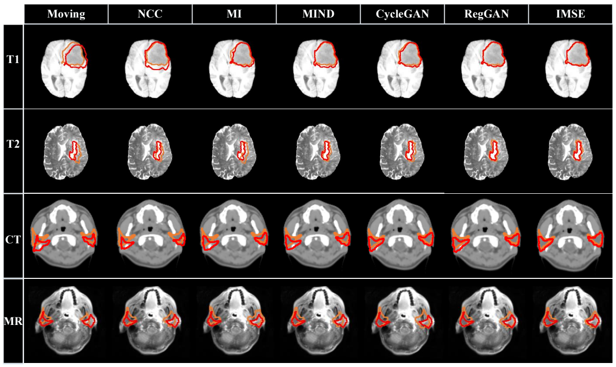

Figure 5. Registration results for various registration methods. Four rows correspond to T1→T2, T2→T1, CT→MR and MR→CT registrations, respectively. Orange contours are based on T1 and CT images. Red contours are based on T2 and MR images. The first column (Moving) shows the contour difference without registration.

T2 → T1	N=2	N=[2,10]	N=[2,30]	N=[2,50]
Dice	0.85	0.88	**0.89**	**0.89**
HD95	1.21	0.93	0.88	**0.85**
$\|\nabla\phi\|_2$	0.0028	0.0024	**0.0018**	0.0022

MR → CT	N=2	N=[2,10]	N=[2,30]	N=[2,50]
Dice	0.51	0.55	0.57	**0.60**
HD95	8.83	8.21	7.49	**7.26**
$\|\nabla\phi\|_2$	0.020	0.0029	0.0025	**0.0020**

Table 3. In 2D case, shuffle remap adopts different parameter N

two datasets, the source data used to train the estimator were T1 and CT, respectively. The network structure adopted by the evaluator is ResNet [12], which is consistent with the generator used in CycleGAN and RegGAN. We added random affine and non-affine transformations to the moving and target images during training and testing, including angular rotations of [-3,3], displacement of [-8%,+ 8%], scaling of [8%,+8%]. The non-affine transformation was generated by spatially transforming the moving and target images using elastic transformations followed by Gaussian smoothing. The degree of deformation was 80 and the radius of Gaussian smoothing was 12.

Registration results of various methods based on T1-T2 and MR-CT datasets were summarized in Table 2. We performed both forward and reverse registration for each dataset. Registration performance was measured using Dice, Hausdorff distances, and the smoothness which was defined by the average gradient of the deformation field (in the case of 2D, we only counted the slices that contained contours). For T1-T2 and MR-CT datasets, IMSE achieved the best results based on all metrics, both in 2D and 3D. Because the registration network used by all methods was the same, the influence of the model was excluded. IMSE had both high registration accuracy and a smoother deformation field, which in general is difficult to achieve simultaneously. It was likely due to the accurate estimate of spatial errors with the adoption of the estimator, which allowed the registration model to achieve a better alignment at a small deformation cost. In addition, our results show that when the Bézier curve was used for data enhancement, it has no performance advantage compared with other methods. But the combination of IMSE and Shuffle Remap provided the best performance among all registration methods. We want to further explore how different segments N of shuffle remap affect the registration results. As shown in Table 3, we can first see that with the increase of N, the performance will improve. In addition, for T1-T2, there is no significant difference between N of 30 and 50. For MR-CT, a larger N obviously brings better scores since MR-CT shows larger distribution differences than T1-T2, which requires more complex style enhancement.

4.3. IMSE for Traditional Registration

Compared to the existing deep learning registration methods which directly provide deformation fields, IMSE essentially evaluates the similarity between multi-modal images through neural networks. Then, the neural network can accurately achieve backward propagation. Therefore, IMSE can be readily integrated into the traditional registration process, such as replacing the L_{sim} function in Eq 1 with a trained evaluator. By first initializing a deformation field of 0 and then optimizing it through similarity loss (Similarity operator or IMSE) and regularization loss(Eq 7). We set the deformation field size to [64,128,128], the learning rate to 1, the number of iterations to 200, and the optimizer to $adam$. In this section, we not only evaluate the traditional multi-modal registration, but also compare IMSE with single-

Moving→Target	Methods	Dice ↑	HD95 ↓	$\|\nabla\phi\|_2$ ↓
T2 → T1	NCC	0.84 ± 0.03	3.40± 1.14	0.016
	MI	0.85 ± 0.03	3.26± 1.22	0.06
	MIND	0.89 ± 0.02	3.01± 1.16	**0.002**
	IMSE	**0.91± 0.01**	**2.62± 0.81**	**0.002**
MR → CT	NCC	0.54 ± 0.03	5.18± 1.20	0.011
	MI	0.55 ± 0.02	5.29± 1.18	0.05
	MIND	0.55 ± 0.02	4.93± 1.24	0.008
	IMSE	**0.61± 0.01**	**4.37± 0.81**	**0.007**
T1 → T1	MAE	0.64 ± 0.03	8.38± 1.43	0.018
	IMSE	**0.67 ± 0.02**	**7.4± 1.22**	**0.005**
CT → CT	MAE	0.56 ± 0.02	4.92± 1.07	0.012
	IMSE	**0.59 ± 0.02**	**4.80±0.87**	**0.005**

Table 4. Comparison of registration results for traditional algorithms.

modal registration using MAE as an optimization measure. Single-modal registration is evaluated using T1→T1 and CT→CT data. Since the single-modal images were from different patients with spatial differences significantly larger than those from the same patient, we screened 5 patients whose spatial location differences were relatively small to mimic the scenario of registering images from the same patient. The results are shown in Table 4. As the metric for registration optimization, IMSE still achieved the best performance in multi-modal conditions. We focus on the results of single-modal registration. IMSE performed much better than MAE in all aspects. This is because even in single-modal datasets, there are still residual distribution differences. T1-T1 or CT-CT should be categorized as relatively single-modal data, especially when images from different patients are registered to each other. MAE cannot ignore residual distribution differences whereas IMSE can.

4.4. IMSE for Image-to-image Translation

To explain how IMSE may enable a new paradigm of image-to-image translation, we use multi-modal images x and y as an example where x is the reference image and y is the image awaiting to be translated. Input x and y into IMSE, and IMSE will use x as the reference image to predict the absolute single-modal error:$E(x, y) \approx x - x^{'}$. Where, the distribution of x and $x^{'}$ is consistent. By subtracting $E(x, y)$ from the reference image x, we can get the translated image $x^{'}$, i.e., $x^{'} \approx x - E(x, y)$. In the new paradigm, image-to-image translation is achieved by reverse inference based on the prediction of the evaluator. In Figure 6, we show a few examples of using IMSE to perform image-to-image translation. $x^{'}$ is the modal translated image of y as it has the spatial characteristics of y but the modal characteristics of x. For example, the dental artifact in the reference CT image in Figure 6 remains in the translated image. This also verifies that IMSE is an instance mapping method based on absolute single-modal.

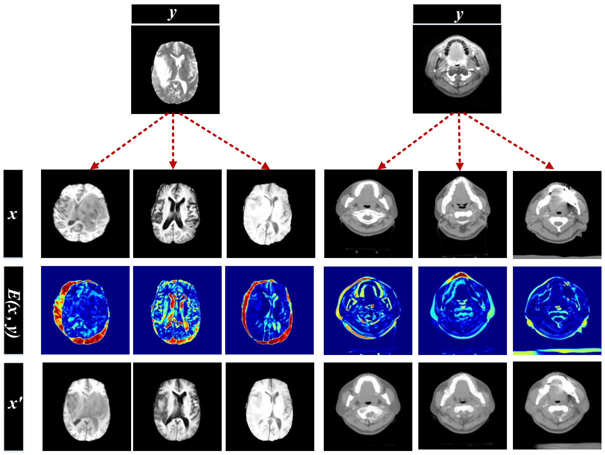

Figure 6. IMSE is used for image to image translation. Where, y is the source image, x is different reference images, $E(x, y)$ is the prediction result of IMSE and $x^{'} = x - E(x, y)$.

Source→Target	Methods	NMAE ↓	PSNR ↑	SSIM ↑
T2 → T1	CycleGAN	0.088	24.1	0.89
	RegGAN	0.071	25.5	0.90
	IMSE	**0.029**	**91.6**	**0.96**
MR → CT	CycleGAN	0.049	22.9	0.88
	RegGAN	0.041	24.1	0.89
	IMSE	**0.022**	**41.3**	**0.93**

Table 5. The results of image-to-image translation.

Table 5 shows the comparison results between IMSE and baseline methods. Spatial transformations were added to x and y in IMSE to prevent alignment between input images. Despite this, IMSE still produces superior results compared to other methods due to the use of an additional target reference image. Thus, the comparison between IMSE and GAN-based methods seems unjust. We aim to investigate the contrasting use scenarios of these two methods, further.

IMSE based image-to-image translation differs from GAN based translation in two aspects. **1)** The result of image-to-image translation based on GAN is unique whereas IMSE based depends on the characteristics of the reference image. **2)** GAN based image-to-image translation requires two modal data whereas IMSE based translation only uses one modal data for training. IMSE is very promising for

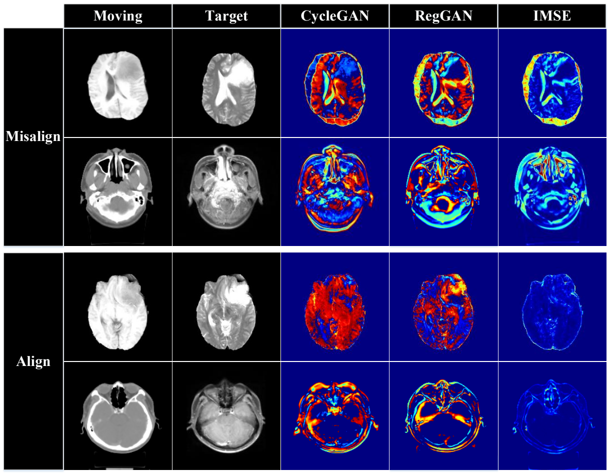

Figure 7. Demonstration of errors estimated by the CycleGAN, RegGAN and IMSE methods in both misaligned and aligned cases.

medical image-to-image translation. It requires less data for training and therefore can reduce expensive data costs. It is better suited to the complexity and diversity of medical scenes. Also, medical image translation requires high accuracy. It is always beneficial to ensure that the characteristics of the translated image come from the desired reference image.

4.5. IMSE for Spatial Error

Without labels, accurate evaluation of registration results is always challenging. IMSE has great potential in offering an objective metric to accurately evaluate registration performance. Since the output of IMSE has the same size as the input image, it can provide pixel-wise registration error estimation. In Figure 7, we demonstrate spatial errors estimated by the CycleGAN, RegGAN and IMSE methods in both misaligned and aligned cases. CycleGAN and RegGAN translated the moving image first and then calculate MAE between the translated image and the target image.

The estimated spatial errors in the CycleGAN and RegGAN methods could not exclude the effect of residual distribution between the translated and target images. The errors remained significant even if images were well aligned. As a comparison, IMSE reported large spatial errors in misaligned regions but small spatial errors in aligned regions.

To quantitatively demonstrate the potential of IMSE for registration performance evaluation, we explored the correlation between IMSE output and Dice. Based on the joint region of Moving and Target masks, we calculated the mean value of IMSE outputs. Then we subtracted the mean absolute value from 1 and performed a normalization to get the value which indicated registration performance for IMSE. Similarly, we calculated values indicating registration performance for CycleGAN and RegGAN. The tests were performed in 3D due to limited test data. To increase the number of test data, we simulated more test data through many random transformations of the image. Figure 8 clearly demonstrates the positive correlation between IMSE and Dice, which confirms the potential of using IMSE to accurately evaluate spatial errors. As a comparison, CycleGAN and RegGAN do not have obvious correlation with Dice, especially on the T1-T2 dataset.

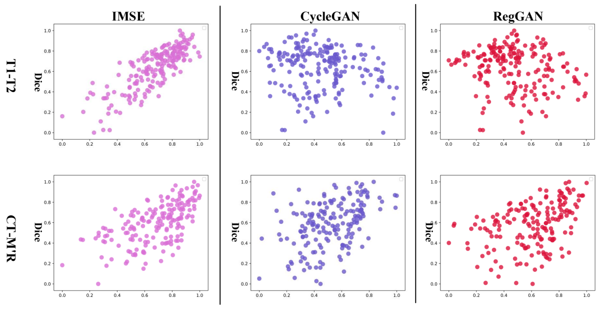

Figure 8. Correlation of estimated spatial errors with Dice for the IMSE, CycleGAN, RegGAN methods.

5. Discussion

In this study, we propose a new approach IMSE for multimodal image registration. IMSE is simple yet powerful. As a metric, IMSE can be used to evaluate the registration results or be combined with traditional registration process. It can also be used to perform image-to-image translation. IMSE uses neural networks instead of similarity operators as loss functions to achieve better results. It also stimulates thoughts of using neural network as indirect constraints to solve challenging problems, instead of committing significant time and efforts searching for operators or loss functions for better performance. Shuffle Remap is an essential component of IMSE. It greatly reduces the amount of data required for model training. All trained models in the current study used only one modal of data and the evaluator has the capability to evaluate and translation unseen data as well. Even though our study focused on medical images, the principle of IMSE should apply to natural images as well. In the future, we will continue investigating IMSE from three aspects. **1)** We have demonstrated the correlation between IMSE and Dice. But the correlation is not as strong as we prefer. We will explore various ways of calculating spatial errors to see if it is possible to improve correlation. **2)** The evaluator is based on neural networks. If it is integrated into the traditional algorithm, backward propagation requires more time. We will explore whether the evaluator can achieve similar results using a simpler architecture. **3)** We will explore other variants of Shuffle Remap and prove that Shuffle Remap as a method of style enhancement can be more impactful in domain generalization.

References

[1] 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. 2

[2] 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. 3, 5

[3] Thuvanan Borvornvitchotikarn and Werasak Kurutach. mirid: Multi-modal image registration using modality-independent and rotation-invariant descriptor. *Symmetry*, 12(12):2078, 2020. 2, 3

[4] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pages 301–318. Springer, 2020. 3

[5] Bob D de Vos, Bas HM van der Velden, Jörg Sander, Kenneth GA Gilhuijs, Marius Staring, and Ivana Išgum. Mutual information for unsupervised deep learning image registration. In *Medical Imaging 2020: Image Processing*, volume 11313, pages 155–161. SPIE, 2020. 2, 3

[6] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, pages 200–216. Springer, 2020. 3

[7] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8208–8217, 2021. 3

[8] Yabo Fu, Yang Lei, Tonghe Wang, Jun Zhou, Walter J Curran, Pretesh Patel, Tian Liu, and Xiaofeng Yang. Deformable mri-ct liver image registration using convolutional neural network with modality independent neighborhood descriptors. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 102–107. SPIE, 2021. 2, 3

[9] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 3

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 3

[11] Courtney K Guo. *Multi-modal image registration with unsupervised deep learning*. PhD thesis, Massachusetts Institute of Technology, 2019. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[13] 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, 3, 5

[14] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020. 3

[15] Chongfei Huang, Chenhui Qiu, Zhiyi Peng, Jing Yuan, and Dexing Kong. Iterative reweighted local cross correlation method for nonlinear registration of multiphase liver ct images. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 136–140. IEEE, 2021. 2, 3

[16] 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

[17] 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*, 34:1964–1978, 2021. 2, 5

[18] 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*, 34:1964–1978, 2021. 3

[19] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 3

[20] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 3

[21] Chenyu Lian, Xiaomeng Li, Lingke Kong, Jiacheng Wang, Wei Zhang, Xiaoyang Huang, and Liansheng Wang. Cocyclereg: Collaborative cycle-consistency method for multimodal medical image registration. *Neurocomputing*, 2022. 2

[22] Shan Liu, Bo Yang, Yang Wang, Jiawei Tian, Lirong Yin, and Wenfeng Zheng. 2d/3d multimode medical image registration based on normalized cross-correlation. *Applied Sciences*, 12(6):2828, 2022. 2, 3, 5

[23] 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, 3

[24] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 5

[25] Michael E Mortenson. *Mathematics for computer graphics applications*. Industrial Press Inc., 1999. 5

[26] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities

via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 3

[27] Josien PW Pluim, JB Antoine Maintz, and Max A Viergever. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8):986–1004, 2003. 2, 3, 5

[28] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 3

[29] 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

[30] 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. 3

[31] Y Raghavender Rao, Nikhil Prathapani, and E Nagabhooshanam. Application of normalized cross correlation to image registration. *International Journal of Research in Engineering and Technology*, 3(5):12–16, 2014. 2, 3

[32] Jignesh N Sarvaiya, Suprava Patnaik, and Salman Bombaywala. Image registration by template matching using normalized cross-correlation. In *2009 international conference on advances in computing, control, and telecommunication technologies*, pages 819–822. IEEE, 2009. 2, 3

[33] Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, page 109115, 2022. 3

[34] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *European Conference on Computer Vision*, pages 68–83. Springer, 2020. 3

[35] Gerard Snaauw, Michele Sasdelli, Gabriel Maicas, Stephan Lau, Johan Verjans, Mark Jenkinson, and Gustavo Carneiro. Mutual information neural estimation for unsupervised multi-modal registration of brain images. *arXiv preprint arXiv:2201.10305*, 2022. 2, 3

[36] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging*, 39(12):4237–4248, 2020. 3

[37] Dongming Wei, Sahar Ahmad, Jiayu Huo, Wen Peng, Yunhao Ge, Zhong Xue, Pew-Thian Yap, Wentao Li, Dinggang Shen, and Qian Wang. Synthesis and inpainting-based mr-ct registration for image-guided thermal ablation of liver tumors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 512–520. Springer, 2019. 3

[38] 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

[39] Feng Yang, Mingyue Ding, and Xuming Zhang. Non-rigid multi-modal 3d medical image registration based on foveated modality independent neighborhood descriptor. *Sensors*, 19(21):4675, 2019. 2, 3

[40] Sergey Zakharov, Wadim Kehl, and Slobodan Ilic. Deceptionnet: Network-driven domain randomization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 532–541, 2019. 3

[41] Ziqi Zhou, Lei Qi, Xin Yang, Dong Ni, and Yinghuan Shi. Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20856–20865, 2022. 3

[42] 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, 3, 5