Unsupervised Multi-modal Medical Image Registration via Invertible Translation

Mengjie Guo[1,2]

[1] University of Birmingham, Birmingham, UK
[2] Southern University of Science and Technology, Shenzhen, China
guomj01@gmail.com

Abstract. In medical imaging, the alignment of multi-modal images plays a critical role in providing comprehensive information for image-guided therapies. Despite its importance, multi-modal image registration poses significant challenges due to the complex and often unknown spatial relationships between different image modalities. To address this, we introduce a novel unsupervised translation-based multi-modal registration method, termed Invertible Neural Network-based Registration (INNReg). INNReg consists of an image-to-image translation network that converts multi-modal images into mono-modal counterparts and a registration network that uses the translated mono-modal images to align the multi-modal images. Specifically, to ensure the preservation of geometric consistency after image translation, we introduce an Invertible Neural Network (INN) that leverages a dynamic depthwise convolution-based local attention mechanism. Additionally, we design a novel barrier loss function based on Normalized Mutual Information to impose constraints on the registration network, which enhances the registration accuracy. The superior performance of INNReg is demonstrated through experiments on two public multi-modal medical image datasets, including MRI T1/T2 and MRI/CT pairs. The code is available at https://github.com/MeggieGuo/INNReg.

Keywords: Multi-modal image registration · Invertible Neural Network · Image translation · Barrier function

1 Introduction

Multi-modal image registration aims at aligning images from different modalities into a common coordinate system, facilitating accurate comparison, integration, and analysis. This process is particularly critical in the fusion of medical images from distinct modalities such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Computed Tomography (CT), etc. In medical imaging, each modality presents unique insights into anatomical structures by highlighting various physical attributes, and the fusion across modalities not only leverages the unique strengths of each modality but also provides a more holistic understanding of patient anatomy, significantly enhancing diagnostic accuracy and the effectiveness of treatment strategies.

The challenge of multi-modal image registration lies in bridging the gap between distinct image characteristics inherent to each modality. Early works for addressing this challenge mainly use supervised learning methods that rely on labelled data, such as ground-truth deformation fields or segmentation masks [31]. However, the applicability of these supervised learning methods is severely limited by the scarcity of annotated data due to the heavy effort required for manual labelling. Particularly, for registration, it is difficult to acquire human-labelled deformation fields. Due to these reasons, many research [4,7,24,25] instead focus on unsupervised learning strategies and attempt to design proper qualitative similarity metrics on the alignment between images from different modalities. These methods eliminate the need for labelled data, though their performance heavily relies on the design of similarity metrics, which can be challenging to define.

For instance, metrics like the Sum of Squared Differences (SSD) and Normalized Cross-Correlation (NCC) are unsuitable for multi-modal contexts, often resulting in suboptimal registration outcomes [3, 4]. The Modality-Independent Neighbourhood Descriptor (MIND) focuses on local structural information and exhibits poor performance in global alignment tasks [13]. On the other hand, while Normalized Mutual Information (NMI) fares better in global alignment, it struggles with local alignment tasks [25].

Recent works on multi-modal image registration have introduced a new deep learning-based paradigm. This paradigm typically consists of a Generative Adversarial Network (GAN) [17] that translates images of different modalities into the same modality, alongside a mono-modal image registration network for aligning the translated images. It is noteworthy that within this paradigm, the objective is for the GAN to harmonize the style of different modalities after translation while preserving the underlying geometry. Subsequently, the registration network learns the deformation field and warps the image geometry accordingly. In contrast to the aforementioned unsupervised learning strategies reliant on similarity metrics, translation-based registration methods offer a workaround to the challenge of devising effective metrics. However, they frequently introduce geometric inconsistencies, leading to a notable decline in registration performance and potential mode collapse when employing GANs in the translation process [1].

In this paper, we propose a novel unsupervised translation-based multi-modal registration method, termed Invertible Neural Network-based Registration (INNReg). Unlike most existing translation-based methods that rely on GANs, INNReg utilizes an Invertible Neural Network (INN) as the translation network. The invertible nature of INN facilitates the learning of modality mapping while ensuring geometric consistency. Additionally, to enhance feature extraction during the translation process, we augment INN with a dynamic depthwise convolution (DDWConv) based on a local attention scheme, which encourages the network to focus on specific regions of the source image [12]. Furthermore, we propose a novel *barrier* NMI loss function to impose constraints on the registration network, addressing the issue of low registration accuracy caused by geometric inconsistency in the translation. To assess the effectiveness of our proposed method, we conduct extensive experiments on two public multi-modal

medical image datasets, including MRI T1/T2 and MRI/CT image pairs. Both qualitative and quantitative analyses demonstrate the competitive performance of our INNReg method in multi-modal registration

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

- We propose INNReg, the first multi-modal image registration method that integrates a translation task through INN. Notably, it effectively addresses the inherent issue of geometric inconsistency found in previous multi-modal image registration approaches that incorporate image translation. In these methods, the image after translation alters not only the appearance but also the underlying geometry of the original image.
- We integrate local attention through dynamic depthwise convolution to capture fine-grained details during the translation process. This method further ensures the preservation of geometric consistency during image translation,
- We propose a novel barrier NMI loss function to constrain the registration network, effectively circumventing the inherent limitations of NMI and thereby improving registration accuracy.
- Through extensive experiments, we demonstrate the superior qualitative and quantitative performance of INNReg compared to state-of-the-art multi-modal image registration approaches.

2 Related Work

2.1 I2I Translation-based Multi-modal Image Registration

In recent years, multi-modal image registration has seen the emergence of several unsupervised translation-based methods. These methods generally adhere to a registration-by-translation paradigm, wherein an Image-to-Image (I2I) translation network is tasked with synthesizing images that mimic the target modality's appearance. This allows for the application of mono-modality similarity metrics to multi-model registration tasks. Among unsupervised translation techniques, cycle-consistent generative adversarial networks (CycleGAN [42]) have gained popularity due to their ability to enforce content preservation through cycle consistency. However, this approach often generates multiple solutions, risking the anatomical accuracy of translated images by introducing artefacts and potentially degrading the quality of multi-modal registration [22].

To avoid these challenges, various strategies have been explored beyond the scope of CycleGAN. Qin et al. [29] leveraged image disentanglement to segregate images into domain-invariant shape features and domain-specific appearance features, facilitating the training of a registration network using the shape features across modalities. However, it is challenging to clearly define and constrain domain-invariant and domain-specific features between different modalities. Arar et al. [1] proposed a methodology aiming to make the translation and registration steps commutative, thereby indirectly promoting the structural consistency of the translation network. However, the reliance on a GAN-based framework implies that structural consistency could be compromised by the discriminator's influence.

2.2 Invertible Neural Network

Invertible Neural Network (INN) [10,11] has several advantages over GAN, which are commonly used in conventional forward propagation methods. Firstly, GAN often faces significant information loss, while INN only results in slight information loss benefit from both forward and backward computations. Secondly, INN has a significantly lower risk of mode collapse, which is a common issue in GAN training. These advantages of INN make it particularly suited for applications such as image fusion [40, 41], image denoising [15] and I2I translation [2]. For example, Zhao et al. [40] utilize the INN blocks for lossless information transmission in the encoder of their image fusion model. Ardizzone et al. [2] combine an INN with an unconstrained feed-forward network for conditioning to address the task of diverse I2I translation for natural images. Huang et al. [15] propose a novel wavelet-inspired invertible network with redundant invertible sparsifying transforms for image denoising. The reversible architecture of INNs enables precise manipulation of latent spaces, facilitating complex generative tasks while guaranteeing accurate reconstruction of the original input data. INNs have also shown significant promise in enhancing image quality by establishing a direct correlation between inputs and their high-resolution outputs. Despite these advancements, the potential of INNs in the realm of I2I translation remains untapped.

2.3 Local Attention by Dynamic Depthwise Convolution

Attention mechanisms [35] are commonly-used techniques in the fields of image processing [26, 39] and computer vision [14, 36] for encoding long-range dependency in extracted features. In I2I translation, the integration of attention mechanisms has emerged as a significant advancement, enabling more focused and context-aware transformations between source and target domains. To list a few, Tang et al. [34] proposed an Attention-Guided Generative Adversarial Network (AGGAN) to transfer high-level semantic parts of images to obtain high-quality images; Tang et al. [33] developed a Multi-Channel Attention Selection GAN to follow external semantic guidance for I2I translation.

In vision tasks, attention usually acts as a dynamic information aggregator in spatial and temporal domains. Specifically, the local attention mechanism forms the keys and values in a window that the query lies in. The attention output is the weighted aggregation of the corresponding values in the local window, where the weights are the softmax normalisation of the dot-product between the queries and the keys. A notable work in this category is [12], which extensively analyses local attention, and establishes the connection between dynamic depthwise convolution and local self-attention by connecting the attention weights for self-attention and the dynamic weights for convolution. Moreover, they empirically observe that the local attention models based on dynamic depthwise convolution perform on par with or slightly better than previous implementations but have lower computational complexity.

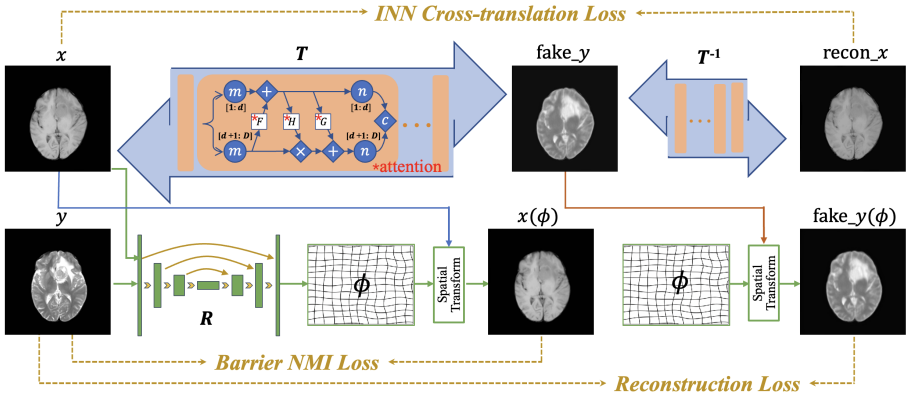

Fig. 1: The framework of the proposed INNReg for multi-modal image registration. In the translation process (**Top**), the forward transformation T translates the source image x into the domain of y as fake_y, and the reverse transformation T^{-1} translates fake_y back into the domain of x as recon_x. The registration process (**Bottom**) first uses the registration network R to predict the deformation field ϕ from x and y, and then uses ϕ and a Spatial Transform function to warp x and fake_y, resulting in $x(\phi)$ and fake_$y(\phi)$, respectively.

3 Methodology

We propose a translation-based multi-modal registration method, referred to as Invertible Neural Network-based Registration (INNReg). INNReg includes an invertible I2I translation network for translating images across various modalities and a deformable registration network for aligning multi-modal medical images based on the translated images. We introduce the registration network in Section 3.1, and describe the invertible translation network in Section 3.2. To enhance feature extraction of the translation network, we equip it with a dynamic depthwise convolution-based local attention scheme. The losses in the proposed INNReg are detailed in Section 3.3. The pipeline of our method is depicted in Figure 1. Note that the registration and translation networks are trained jointly, while only the registration network is used in the test.

3.1 Registration Model

The registration model first employs a network R to learn the deformation field $\phi = R(x, y)$ from a pair of images (x, y) with distinct modalities. The field ϕ is a matrix of 2D vectors, where each vector indicates the displacement for aligning individual pixels from the source image x with the structure of the target image y. With ϕ in hand, we warp x to align with y by a spatial transform function. The learned ϕ is domain-free, meaning that it applies to both the domain $\mathcal{X}$ of the source images and $\mathcal{Y}$ of the target images.

3.2 Translation Model

We aim to design a translation model in INNReg to harmonize image styles from different modalities while preserving the crucial geometry details inherent to each image, which is particularly vital for translation-based multi-modal image registration methods. To achieve this, we develop the translation model based on INN, which allows for mutual generations of features and therefore ensures a high fidelity of information preservation between its input and output.

The INN in our translation model can be described as a function $T : \mathcal{X} \to \mathcal{Y}$ mapping from the source image domain $\mathcal{X}$ to the target image domain $\mathcal{Y}$ and consists of a stack of Invertible Blocks (InvBlocks) $\{T_i\}_{i=1}^{k}$. Specifically, for any source image x, its translation via T is

$$\text{fake\_}y = T(x) = T_0 \circ T_1 \circ T_2 \circ \cdots \circ T_k(x), \tag{1}$$

and the inverse translation back to the source image domain $\mathcal{X}$ is

$$\text{recon\_}x = T^{-1}(\text{fake\_}y) = T_k^{-1} \circ T_{k-1}^{-1} \circ \cdots \circ T_0^{-1}(\text{fake\_}y), \tag{2}$$

where $T^{-1} : \mathcal{Y} \to \mathcal{X}$ is the inverse of T and each T_i^{-1} is the inverse of the InvBlock T_i.

In our method, each InvBlock T_i is realized through an affine coupling layer [11]. Given a D-dimensional vector m and an integer $d \in [1, D)$, the layer first splits m into two vectors $m_{1:d}$ and $m_{d+1:D}$, and then performs the additive affine transformations:

$$n_{1:d} = m_{1:d} + F\left(m_{d+1:D}\right) \tag{3}$$

$$n_{d+1:D} = m_{d+1:D} \odot \exp\left(H\left(n_{1:d}\right)\right) + G\left(n_{1:d}\right) \tag{4}$$

where $\odot$ is the Hadamard product, and F, H, and G are arbitrarily complicated functions of m and not necessarily invertible. In our implementation, the three functions F, H, and G are realized by a common densely connected convolutional block, referred to as DenseBlock in [32,37]. Note that given $n_{1:d}$ and $n_{d+1:D}$, the inverse transformations of (3)–(4) can be computed as

$$m_{d+1:D} = \left(n_{d+1:D} - G\left(n_{1:d}\right)\right) \odot \exp\left(-H\left(n_{1:d}\right)\right) \tag{5}$$

$$m_{1:d} = n_{1:d} - F\left(m_{d+1:D}\right). \tag{6}$$

Local Attention by DDWConv. To improve the feature extraction ability of the translation model, we equip INN with a dynamic depthwise convolution (DDWConv)-based local attention scheme. Local attention mechanisms are particularly useful in I2I translation tasks, in the sense that they can help the model refine the details in the generated output by focusing on specific regions of the input image.

Unlike existing local attention schemes that are mainly designed based on feature aggregation modules with Key, Query, and Value, we adapt DDWConv

to realize local attention. The recent work [12] points out that to yield excellent performance, local attention schemes should satisfy three principles: sparse connectivity, weight sharing, and dynamic weighting. These characteristics also exist in dynamic depthwise convolution in CNN, and compared to common local attention realizations, the computational cost of using DDWConv is often lower. Inspired by this, we realize local attention by DDWConv. To illustrate DDWConv, we first introduce depthwise convolution:

$$DepthwiseConv_{(i,j)} = \sum_{p,q}^{P,Q} W_{(p,q)} \cdot x_{(i+p,j+q)}, \tag{7}$$

where (i, j) is a coordinate, P and Q are the height and width of x, respectively, W represents weight, and x is an input feature map. In DDWConv, the weight W in (7) is learned *dynamically* by aggregating multiple convolution kernels $\pi_k W_k$, $k = 1, \ldots, K$:

$$W(x) = \sum_{k=1}^{K} \pi_k(x) W_k. \tag{8}$$

Note that different from the (static) depthwise convolution in which W is fixed, the weight $W(x)$ in DDWConv varies for each input feature map x. Due to this feature, DDWConv suits the input's characteristics and, therefore, can potentially extract more relevant and discriminative features from the input data, which further leads to the superior performance of local attention in feature extraction. A sketch of the DDWConv-based local attention scheme is provided in Fig. 2.

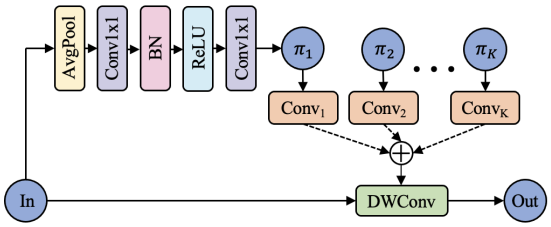

Fig. 2: DDWConv-based local attention.

3.3 Loss Functions

The proposed INNReg is trained in an end-to-end manner. The total loss function consists of four components, including a pixel-wise reconstruction loss, a barrier NMI registration loss, an INN cross-translation loss, and an additional smooth

loss. During the training, we use the barrier NMI registration loss to constrain the registration network, the INN cross-translation loss to optimize the translation network, the pixel-wise reconstruction loss to jointly optimize the registration network and the translation network to generate outputs that closely match the target image, and the smooth loss to preserve the smoothness of the predicted deformation field.

Pixel-wise Reconstruction Loss. The pixel-wise reconstruction loss is the key to the success of the translation-based methods in handling multi-modal registration problems. In our method, it is the ℓ_1-reconstruction loss between the generated results (i.e., fake_$y(\phi)$) and the target image y:

$$\mathcal{L}_{\text{recon}}(T, R) = \left\| \text{fake\_}y(\phi) - y \right\|_1, \tag{9}$$

where T and R are the translation and the registration network, respectively. The use of the loss (9) enforces fake_$y(\phi)$ to approximate the target image y.

Barrier NMI Registration Loss. The barrier NMI registration loss is proposed to improve the accuracy of the registration network. In the translation-based multi-modal registration methods, the deformation in the translation network often deteriorates the accuracy of the registration network. Our idea of addressing this issue is to constrain the similarity between the warped source image and the target image.

In our method, we adopt the widely used Normalized Mutual Information (NMI) to measure the similarity between two images by quantifying their mutual information of intensity values. The NMI metric between the warped source image $x(\phi)$ and the target image y can be expressed as

$$I_{\text{NMI}}(y, x(\phi)) = \frac{H(y) + H(x(\phi))}{M(y, x(\phi))}, \tag{10}$$

where $H(y)$ and $H(x(\phi))$ denote the marginal entropies of the source and the target image, respectively, and $M(y, x(\phi))$ is the mutual information between $x(\phi)$ and y. Although NMI can also be used as a loss function in DL models for multi-modal image registration, it comes with certain disadvantages [28]. In particular, they cannot effectively capture complex or nonlinear relationships in multi-modal data, potentially limiting the performance of the multi-modal registration network.

Based on NMI, we propose the following barrier loss function:

$$\mathcal{L}_{\text{barrier}}(I_{\text{NMI}}, b) = -\log(b - I_{\text{NMI}}), \tag{11}$$

where b is a manually set threshold. The barrier function value tends to infinity when I_{NMI} approaches b and, therefore, adding the barrier loss to the total loss *rigorously* guarantees $I_{\text{NMI}} < b$. This property helps to avoid the accuracy degradation of the registration network caused by deformation in the translation

network. To see this, note that if a severe deformation occurs in the translation network, then the registration network will generate an inaccurate deformation field, which further leads to a huge barrier loss. Additionally, since the barrier function value changes slowly when I_{NMI} is not too close to the threshold b, the influence of this barrier loss diminishes, ensuring that the limitations of NMI do not significantly impact the registration outcome.

INN Cross-Translation Loss. One major challenge in I2I translation tasks is the requirement of paired images. For unpaired data, previous methods typically employ the GAN loss for translation tasks [5, 23], but it often results in mode collapse. In this method, we introduce an INN cross-translation loss as a constraint of our translation model, which takes the following form:

$$\mathcal{L}_{\text{INN}}(T) = \|\text{recon\_}x - x\|_1 \tag{12}$$

where x is the source image, T is the INN translation network, and $\text{recon\_}x = T^{-1} \circ T(x)$ with T^{-1} being the reverse transformation of T. The reversibility of INN ensures that the translation process using (12) is coherent and maintains the critical attributes of the original images, even in the absence of paired training samples. An intuitive explanation is that the INN cross-translation loss preserves the content of the image and only changes the domain-specific attributes.

Smooth Loss. In deformable registration, the deformation field often suffers from the issue of over-distortion, which makes it hard to establish a precise match between two unaligned images. To prevent over-distortion, we use a smooth loss to restrict the gradients of the predicted deformation field ϕ:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{N} \sum_{p} \|\nabla \phi(p)\|_2^2, \tag{13}$$

where each p is a pixel in ϕ and N represents the total number of pixels.

Total Loss. With the above four loss functions, the total loss function for training takes the following form:

$$\mathcal{L}_{\text{Total}} = \lambda_\alpha \mathcal{L}_{\text{recon}} + \lambda_\beta \mathcal{L}_{\text{barrier}} + \lambda_\gamma \mathcal{L}_{\text{INN}} + \lambda_\delta \mathcal{L}_{\text{smooth}} \tag{14}$$

where $\lambda_\alpha, \lambda_\beta, \lambda_\gamma$ and λ_δ are the weights to each loss term. The purpose of this loss is to improve the multi-modal registration performance by letting the translation network focus on the modality mapping and the registration network on geometry warping.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate the registration performance of our INNReg method on two publicly accessible multi-modal medical image datasets, including a T1/T2 weighted MRI image dataset and an MRI/CT image dataset.

Table 1: The multi-modal image registration results on T1/T2 and MRI/CT dataset. The best and the second best results are marked in **bold** and by asterisk*, respectively.

Methods		VM	Arar et al.	RegGAN	Chen et al.	INNReg(Ours)
T1/T2	SSIM(%)↑	89.414	83.813	89.971*	89.386	**90.555**
	NCC(%)↑	98.03	95.309	98.238	98.889*	**98.746**
	Dice(%)↑	86.984*	83.296	84.643	81.274	**87.117**
	HD95↓	128.318	101.806	**100.660**	104.473	101.371*
	smooth(%)↓	0.238	**0.007**	0.232	5.303	0.101*
MRI/CT	SSIM(%)↑	54.74	64.791	71.637*	63.308	**73.067**
	NCC(%)↑	55.903	83.776	88.716*	80.455	**89.889**
	smooth(%)↓	74.815	**0.006**	0.767	9.113	0.179*

The T1/T2 dataset is from the Brain Tumour Segmentation (BraTS) 2023 Challenge [18] containing segmentation labels and is widely used for medical image registration [23,38]. The labels delineated three clinical tumour regions of interest (ROIs) within the brain and were approved by expert neuroradiologists. The dataset consists of 60 scan pairs with labels, where each scan is a 3D volume of $240 \times 240 \times 155$. In our experiments, we randomly divide the 60 pairs into 54/6 for training/testing, and in each volume, we extract the middle 50 slices as our input, where each slice is a 2D image of the size 240×240.

The MRI/CT dataset is from Harward [19], and in our experiment, we use 920 image pairs for training and 92 for testing. Since the multi-modal image pairs in the aforementioned two datasets are all aligned, we need to synthetically generate unaligned images from them to train the network. Specifically, we adopt the B-spline transformation method used in [8,21] to generate unaligned T1/T2 weighted MRI and MRI/CT pairs. All images used in the experiments are resized as 192×192.

Implementation Details. The implementation of our code uses PyTorch 3.9.0 [27], and the experiments were conducted on a single Nvidia A100 GPU. We use the Adam optimizer [20] on a mini-batch of size 3 with the initial learning rate $lr = 0.0002$ and the momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We train our model for 200 epochs and activate linear learning rate decay after every 20 epochs. We implement the registration network by U-NET [30] with residual connections in the encoder and output paths. The number of INN blocks in the translation model is 8. In the total loss function, the parameters $\lambda_\alpha, \lambda_\beta, \lambda_\gamma$ and λ_δ are asigned as $1, 1, 0.1$, and 10, respectively. The boundary b in the barrier NMI loss is set to 5.

Comparison Methods. To demonstrate the superiority of the proposed INNReg method, we compare it with four widely-used multi-modal image registration methods, including the classical deep learning-based method Voxelmorph [3] and three translation-based methods: Arar et al. [1], RegGAN [22], and Chen et al. [6]. In Voxelmorph, we use the local cross-correlation as the similarity loss.

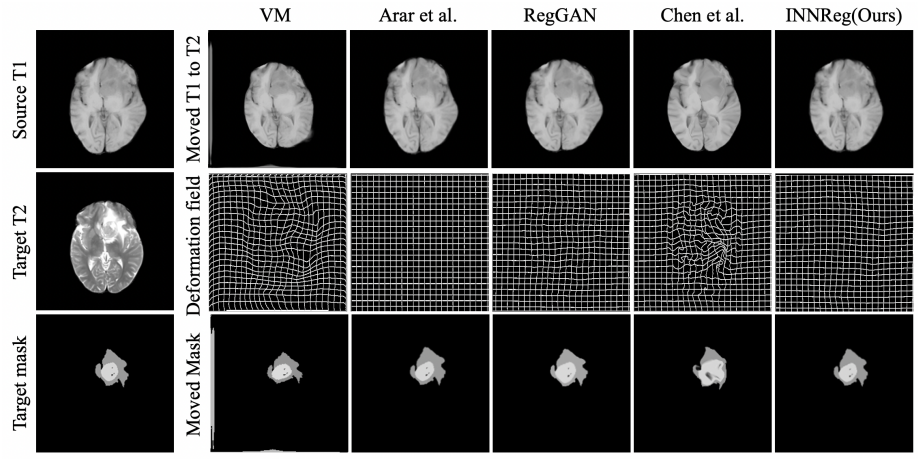

Fig. 3: Visualization results of INNReg against other four methods on T1/T2 images. The first row shows the warped source image $x(\phi)$, and the second displays the corresponding deformation field ϕ. The last row exhibits the warped segmentation mask of the source image.

For a fair comparison, all comparison methods are re-trained using the same training dataset as ours.

Evaluation Metrics. We evaluate the registration performance in terms of image similarity, mask similarity, and diffeomorphism. A higher image and mask similarity and a smoother diffeomorphism indicate higher registration accuracy. To measure image similarity, we adopt two standard metrics, including Structural Similarity (SSIM) and Normalized Correlation Coefficient (NCC), where higher SSIM and NCC indicate higher similarity in image level. Our quantitative metrics for measuring mask similarity include the Dice score [9] and the 95% maximum Hausdorff distance (HD95) [16], which measure the degree of overlap between two regions. The higher Dice score and HD95 depict a higher mask similarity. We measure diffeomorphism by the smoothness of the deformation field ϕ, which can be quantified by the determinant of its Jacobian matrix J_ϕ. Specifically, we count the percentage of the pixels p in each image such that $|J_\phi(\mathbf{p})| \leq 0$, and a smaller percentage implies higher smoothness of the deformation field, which further causes smoother diffeomorphism.

4.2 Results on T1/T2 Dataset

The quantitative registration results on the T1/T2 datasets are summarized in Table 1. Based on these results, we compare the proposed INNReg network with four other methods in terms of five evaluation metrics. From Table 1, we observe that among all the five evaluation metrics, INNReg achieves the best in terms

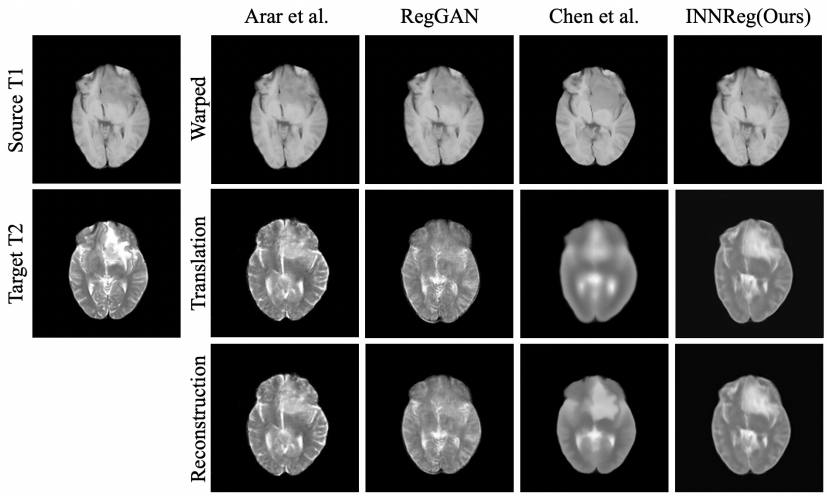

Fig. 4: Visualization results of INNReg against the other three translation-based methods on T1/T2 images. The first row shows the warped source T1 image $x(\phi)$, and the second shows the translated results fake_y. The last row displays the reconstructed results fake_$y(\phi)$.

of three of them (SSIM, NCC, Dice) and the second best in the remaining two (HD95, smooth). Note that although Arar et al. achieves the best smoothness, it is caused by the deformation that occurs in its translation network rather than the anticipated registration network, leading to a negligible deformation field. Consequently, INNReg outperforms the other comparison methods.

Fig. 3 visualizes the multi-modal image registration results on the T1/T2 dataset. Specifically, it illustrates an instance of the warped source image, the corresponding deformation fields, and the warped source segmentation mask. By comparing the warped source images and the target image, we observe that INNReg achieves superior alignment between the source and target images. Furthermore, INNReg generates a smooth deformation field and the most accurate mask movement towards the target.

Fig. 4 displays not only the final warped image but also the intermediate results, which include the translated image (fake_y in Fig. 1) and the reconstruction result (fake_$y(\phi)$ in Fig. 1). From Fig. 4, we find a promising reason for the excellent registration performance of INNReg over the other translation-based methods – both Arar et al and RegGAN exhibit significant deformation in the translated results, which notably impacts registration accuracy; The method proposed by Chen et al. does not have this issue but the quality of the translated image is unsatisfactory. Compared to these three methods, the translated image generated by our INNReg avoids both issues, and the superior translation results further lead to high-quality registration results.

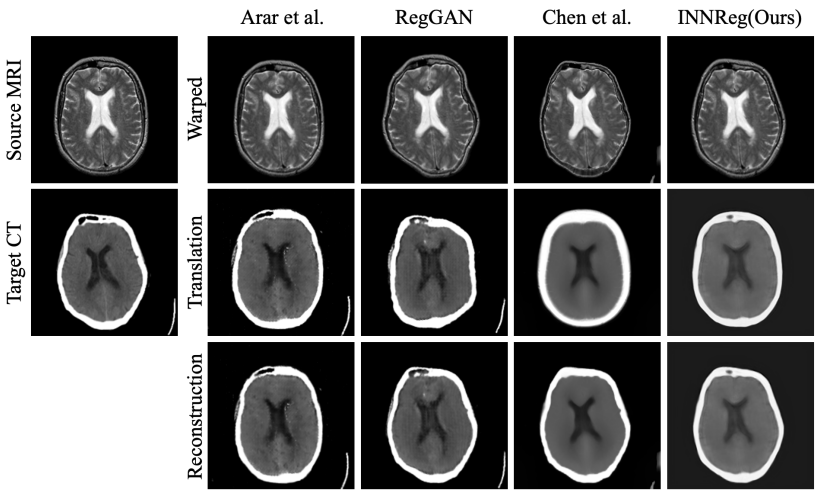

Fig. 5: Visualization results of INNReg against the other three translation-based methods on MRI/CT images. The first row depicts the warped source MRI image $x(\phi)$, and the second shows the translated results fake_y. The last row shows the reconstructed results fake_$y(\phi)$.

4.3 Results on MRI/CT Dataset

We also evaluate the effectiveness of the proposed INNReg on MRI/CT images. Due to the absence of segmentation masks, our evaluation focuses on the image similarity metrics SSIM and NCC, and the diffeomorphism of the deformation field. The quantitative results are displayed in Table 1 and reveal that our method surpasses other compared methods in terms of SSIM and NCC. The smoothness of INNReg is inferior only to Arar et al., due to the negligible deformation in the translation process of Arar et al. as previously discussed.

Fig. 5 visualizes the warped source image, the translated image, and the reconstructed image in the experiments. It shows that INNReg achieves high-accuracy image registration on the MRI/CT image pairs. Additionally, we observe that our translation result remains highly geometry consistent with the source image, and the reconstructed image closely resembles the target image.

4.4 Ablation Study

In this subsection, we perform a series of ablation experiments to examine the impact of the INN cross-translation loss, the barrier NMI loss, and the DDWConv-based local attention on the registration accuracy of the proposed INNReg. These experiments are conducted using T1/T2 images, and the results are summarised in Table 2, in which the Base model refers to the version of INNReg stripped of the INN cross-translation loss, the barrier NMI loss, and DDWConv-based local attention. We evaluate the effectiveness of each component by comparing

Table 2: Ablation experiment results on T1/T2 images. **Bold** indicates the best value.

Methods	SSIM(%)↑	NCC(%)↑	Dice(%)↑	HD95↓	smooth(%)↓
Base(A)	87.621	98.192	83.026	139.55722	0.128
A+Cross translation(B)	88.201	98.379	83.469	139.01316	0.115
AB+NMI(C)	86.191	97.829	82.153	137.67901	0.135
AB+Barrier NMI(D)	**90.84**	98.726	86.84	136.84976	**0.095**
ABD+Attention(Ours)	90.555	**98.746**	**87.117**	**101.37179**	0.100

the performance of the models with and without it. From Table 2, we make the following observations: First, the cross-translation and the barrier NMI loss enhance the registration results in all five metrics; Second, the barrier NMI loss is much more effective compared to NMI in terms of all the metrics; Third, although the DDWConv-based attention slightly deteriorates SSIM and smooth, it significantly improves the segmentation metrics, including Dice and HD95.

5 Conclusion

In this study, we introduced INNReg, an innovative unsupervised translation-based method for multi-modal image registration. Central to our approach is the exploitation of the Invertible Neural Network's (INN) reversible nature, enabling precise modality mapping and geometric consistency in I2I translation. By integrating a dynamic depthwise convolution-based local attention mechanism, our method effectively enhances feature extraction during the translation process. Furthermore, we proposed a novel barrier NMI loss function to rigorously constrain the registration process, which avoids the accuracy degradation of the registration network caused by severe deformation in the translation network. The efficiency and rationality of INNReg were validated across two public multi-modal medical image datasets, including T1/T2 MRI and MRI/CT image pairs. The experiment results demonstrated that INNReg well preserves geometry in translation and achieves high registration accuracy. In the future, we will explore further potentials of INNReg in multi-modal image translation tasks.

References

1. Arar, M., Ginger, Y., Danon, D., Bermano, A.H., Cohen-Or, D.: 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. pp. 13410–13419 (2020)
2. Ardizzone, L., Kruse, J., Lüth, C., Bracher, N., Rother, C., Köthe, U.: Conditional invertible neural networks for diverse image-to-image translation. In: Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42. pp. 373–387. Springer (2021)

3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
4. Cao, X., Yang, J., Wang, L., Xue, Z., Wang, Q., Shen, D.: Deep learning based inter-modality image registration supervised by intra-modality similarity. In: Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9. pp. 55–63. Springer (2018)
5. Chen, Z., Wei, J., Li, R.: Unsupervised multi-modal medical image registration via discriminator-free image-to-image translation. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22) (2022)
6. Chen, Z., Wei, J., Li, R.: Unsupervised multi-modal medical image registration via discriminator-free image-to-image translation. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 834–840. International Joint Conferences on Artificial Intelligence Organization (7 2022). https://doi.org/10.24963/ijcai.2022/117, https://doi.org/10.24963/ijcai.2022/117, main Track
7. De Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I.: 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: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3. pp. 204–212. Springer (2017)
8. Deng, X., Liu, E., Li, S., Duan, Y., Xu, M.: Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. IEEE Transactions on Image Processing **32**, 1078–1091 (2023)
9. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
10. Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516 (2014)
11. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. In: International Conference on Learning Representations (2016)
12. Han, Q., Fan, Z., Dai, Q., Sun, L., Cheng, M.M., Liu, J., Wang, J.: On the connection between local attention and dynamic depth-wise convolution. arXiv preprint arXiv:2106.04263 (2021)
13. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical image analysis **16**(7), 1423–1435 (2012)
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
15. Huang, J.J., Dragotti, P.L.: Winnet: Wavelet-inspired invertible network for image denoising. IEEE Transactions on Image Processing **31**, 4377–4392 (2022)
16. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (1993)
17. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
18. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., et al.: The brain tumor segmentation

(brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). ArXiv (2023)
19. Keith A. Johnson, J.A.B.: The whole brain atlas. `Harvardmedicalwebsite.http://www.med.harvard.edu/AANLIB/home.html`
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Klein, S.: Optimisation methods for medical image registration. Ph.D. thesis, University Utrecht (2008)
22. Kong, L., Lian, C., Huang, D., Hu, Y., Zhou, Q., et al.: Breaking the dilemma of medical image-to-image translation. Advances in Neural Information Processing Systems **34**, 1964–1978 (2021)
23. Liu, Y., Wang, W., Li, Y., Lai, H., Huang, S., Yang, X.: Geometry-consistent adversarial registration model for unsupervised multi-modal medical image registration. IEEE Journal of Biomedical and Health Informatics (2023)
24. Mahapatra, D., Antony, B., Sedai, S., Garnavi, R.: Deformable medical image registration using generative adversarial networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1449–1453. IEEE (2018)
25. Mahapatra, D., Ge, Z., Sedai, S., Chakravorty, R.: Joint registration and segmentation of xray images using generative adversarial networks. In: Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9. pp. 73–80. Springer (2018)
26. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International conference on machine learning. pp. 4055–4064. PMLR (2018)
27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
28. Pluim, J.P., Maintz, J.A., Viergever, M.A.: Mutual-information-based registration of medical images: a survey. IEEE transactions on medical imaging **22**(8), 986–1004 (2003)
29. Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A.: Unsupervised deformable registration for multi-modal images via disentangled representations. In: International Conference on Information Processing in Medical Imaging. pp. 249–261. Springer (2019)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
31. Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N.: A deep metric for multimodal registration. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19. pp. 10–18. Springer (2016)
32. Song, D., Xu, C., Jia, X., Chen, Y., Xu, C., Wang, Y.: Efficient residual dense block search for image super-resolution. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12007–12014 (2020)
33. Tang, H., Torr, P.H., Sebe, N.: Multi-channel attention selection gans for guided image-to-image translation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 6055–6071 (2022)

34. Tang, H., Xu, D., Sebe, N., Yan, Y.: Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
36. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
37. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
38. Xu, H., Yuan, J., Ma, J.: Murf: Mutually reinforcing multi-modal image registration and fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
39. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning. pp. 7354–7363. PMLR (2019)
40. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5906–5916 (2023)
41. Zhou, M., Huang, J., Fu, X., Zhao, F., Hong, D.: Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)
42. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)