DRT: Deformable Region-Based Transformer for Nonrigid Medical Image Registration With a Constraint of Orientation

Sheng Lan, Xiu Li, *Member, IEEE*, and Zhenhua Guo

***Abstract*— Nonrigid medical image registration is a key technology for clinical disease diagnosis by integrating the vital information taken from input images acquired using multiple imaging sensors. Recently, transformer networks have attracted attention in image registration, but splitting images into patches remains a problem. Although there are usually multiple structural regions with different sizes and shapes in medical images, current transformer methods usually use an equal-sized or rectangular patch embedding that cannot represent the structural differences and directionality of different regions. In addition, it may also destroy the semantic consistency of objects. To address this problem, we propose a new deformable region-based structural relevance embedding (DRSRE) module which learns to adaptively split the images into deformable structural regions of different sizes and shapes with orientation constraints rather than using equal-sized or rectangular patches. In this way, our method can represent the structural differences and directionality between regions and effectively preserve the semantic consistency in regions to calculate accurate transformation relationships for nonrigid image registration. We term the DRSRE module with a transformer part as a deformable region-based transformer (DRT) and conduct extensive evaluations of DRT in the context of nonrigid medical image registration, and the results validate that the proposed DRT outperforms the state-of-the-art nonrigid registration methods, achieving 0.748 and 0.886 Dice scores with 0.217% and 0.308% foldings for deformation field on the LPBA40 and ADNI datasets.**

***Index Terms*— Deformable region-based transformer (DRT), medical image, nonrigid image registration, orientation field.**

I. Introduction

NOWADAYS, medical images collected by multiple sensors are considered by clinicians as an important basis to diagnose critical diseases. For a few different images for quantitative analysis, nonrigid registration is an important step to establish the geometric correspondences between the normal and pathological images. Various methods have been proposed for nonrigid medical image registration, which can be divided into traditional methods and learning-based methods. Although many traditional nonrigid registration methods achieve good results for some specific types of medical images, they usually rely on hand-engineered features for generating the deformation field. As no hand-engineered features can be general enough to be suitable for all types of images, learning-based nonrigid registration methods become popular in recent years. Progresses have been made by applying the convolutional neural network (CNN) for image registration, which outperformed most traditional methods. However, the CNN-based methods cannot well represent long-range spatial relations in patches due to the limitations of convolution operations, which reduces the accuracy of nonrigid registration. Recently, transformer networks have achieved success in many computer vision tasks, such as image segmentation [1], [2], image classification [3], [4], image synthesis [5], [6], and object localization [7], [8]. Compared to CNNs, transformers can represent long-range spatial relations in patches, and the features contain more semantic information. In this background, the transformer has also attracted attention in image registration and trials have been made for achieving transformer-based nonrigid medical image registration [15], [16]. However, there still exists challenges in applying them for nonrigid medical image registration.

On the one hand, most of the existing transformers use an equal-sized rectangular patch embedding based on the assumption that an equal-sized rectangular image split is suitable for all images. As there are usually multiple structural regions with different sizes and shapes in medical images, such an equal-sized rectangular patch split method may pose two problems: 1) disregarding the structural differences and directionality of different regions in an image and 2) semantic inconsistency across patches. Objects in different regions may have different shapes. The equal-sized rectangular split patches will potentially destroy the semantic inconsistency of objects, leading to poor registration performance. Fig. 1(a) illustrates two 3-D brain images before and after applying equal-sized patch embedding in a vanilla transformer. Several internal structures in the moving image can be observed to be improperly split. On the other hand, several methods of hierarchical patches and deformable patches have been proposed to overcome the limitations of equal-sized and rectangular patches, respectively. Among methods of hierarchical patches, the most representative works are pyramid vision transformer (PVT) networks [17] and Swin transformer networks [18], they both produce a hierarchical feature representation and split an image into patches with different sizes. However, they still use rectangular patches without considering the

Manuscript received 24 January 2023; revised 20 March 2023; accepted 13 April 2023. Date of publication 8 May 2023; date of current version 19 May 2023. The Associate Editor coordinating the review process was Dr. Jing Lei. *(Corresponding author: Sheng Lan.)*

Sheng Lan is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: q133589976@163.com).

Xiu Li and Zhenhua Guo are with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: li.xiu@sz.tsinghua.edu.cn; zhenhua.guo@sz.tsinghua.edu.cn).

Digital Object Identifier 10.1109/TIM.2023.3273678

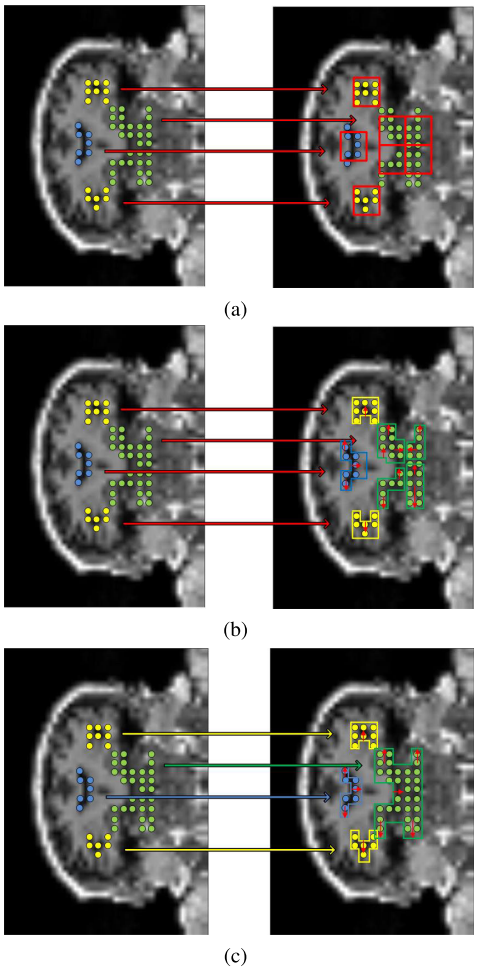


Fig. 1. Example of the DRSRE module. The structural regions with different sizes are marked by boxes with different colors, and offsets of the region boundaries are marked by red arrows. DRSRE can adaptively embed an image into structural regions of different sizes and shapes according to the structural features with orientation constraints, which avoids improper splitting of structures in medical images. (a) Equal-sized rectangular patch split. (b) Deformable patch split [13]. (c) DRSRE.

various shapes of different structures in an image and cannot overcome the limitation of rectangular patches. For methods of deformable patches, the most representative works are deformable convolution (DCN) networks [13], and deformable patch-based transformer (DPT) [14]. Both the DCN and DPT adopt the method of learning the offset of pixels in patches, and adaptively deform the patches in the CNN and transformer, respectively. However, these deformable patch-based methods still learn the offset based on equal-sized split patches, therefore they are still unable to overcome the limitation of equal-sized patches and cannot reflect structural differences in the size of different regions in an image. In addition, they do not consider the structural directionality in regions. As shown in Fig. 1(b), the offset directions of different patches within the same structure can be conflicting or even in opposition.

To address these problems, we propose a new module for image splitting, called deformable region-based structural relevance embedding (DRSRE). Compared with the equal-sized or rectangular patch embedding in existing methods, DRSRE can learn to adaptively split an image into deformable structural regions with different sizes and shapes according to the structural features with orientation constraints. By overcoming the limitations of both the equal-sized and rectangular patch split, DRSRE can represent the structural differences and directionality between regions and effectively preserve the semantic consistency in regions to obtain accurate transformation relationships for nonrigid image registration. To achieve this, we first learn the kernel size of different regions according to the structural similarity between regions and then adjust the shape of the regions according to the orientation fields of the region boundaries as illustrated in Fig. 2. The proposed module can work as a plug-and-play module that can be easily incorporated into transformer architectures. A transformer with a DRSRE module is named deformable region-based transformer (DRT) in this work (see Fig. 3). In DRT, we integrate the DRSRE module into a transformer part to obtain the deformation field for nonrigid medical image registration. With deformable structural regions replacing the equal-sized or rectangular patches, DRT generates complete, discriminative, and authentic features for input images. As there are usually multiple structural regions with different sizes and shapes in medical images, it can achieve high performance in nonrigid registration tasks. To verify the efficiency of DRT, we conduct extensive evaluations of DRT for nonrigid medical image registration. Our method achieves 0.748 and 0.886 Dice scores with 0.217% and 0.308% foldings for the deformation field on the LPBA40 and ADNI datasets which outperforms the state-of-the-art methods.

In summary, the main contributions of this work are as follows.

1) We propose a new DRSRE module. Compared with the equal-sized or rectangular patch embedding in existing methods, DRSRE can learn to adaptively split an image into deformable structural regions of different sizes and shapes. By overcoming the limitations of both the equal-sized and rectangular patch split, DRSRE can represent the structural differences and directionality between regions and effectively preserve the semantic consistency in regions to obtain accurate transformation relationships for nonrigid image registration.
2) Our DRSRE can work as a plug-and-play module integrated into transformers, leading to a DRT in this work. In DRT, we integrate the DRSRE module into a transformer part to obtain the deformation field for nonrigid medical image registration. Specifically, we propose a structural region-attention module in the transformer part to calculate the structural deformation relationship and apply a progressive searching strategy to improve the efficiency of the algorithm.
3) As there are usually multiple structural regions with different sizes and shapes in medical images, we verify DRT in the context of nonrigid medical image registration. The results validate that the proposed DRT outperforms the state-of-the-art nonrigid registration methods.

The remainder of this article is organized as follows. Section II introduces the related transformer methods and deformable patch split methods. Section III discusses the

proposed DRT algorithm in detail. Section IV presents the experimental results with discussions. Finally, Section V concludes our work and discusses possible future works.

II. Related Work

A. Transformer Methods

At present, the transformer network shows a stronger feature expression ability than the CNN, especially for the long-range spatial relations in an image. The first transformer method [9] was proposed for natural language processing. It used attention modules to replace convolutions and obtained state-of-the-art results on the translation task. The attention function was described as mapping a query and a set of key-value pairs to an output, where the queries, keys, and values were packed into matrices Q, K, and V, respectively, as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

Here $\sqrt{d_k}$ is a scaling factor. Dosovitskiy et al. [10] introduced the transformer for image recognition. They proposed a vision transformer (ViT) Network and concluded that an image can be represented by 16×16 feature points. The ViT network consists of an embedding module and alternating layers of multihead self-attention (MSA), layer-norm (LN), and multilayer perceptron (MLP) modules. Given an image i with a size of $H \times W \times C$, it is first reshaped into N patches, and embedded into a sequence of vectors. Then, the vectors are transferred through the MSA, LN, and MLP layers.

After the success of ViT, progress has been made in applying the transformer to various computer vision tasks. For example, Chen et al. [15] developed ViT-V-Net by applying ViT to image registration. To overcome the limitations of fix-sized patches, Wang et al. [17] incorporated the pyramid structure into ViT and proposed the PVT for dense prediction, Liu et al. [18] replace the standard MSA module in a transformer block by a module based on shifted windows and proposed the Swin transformer for image classification and dense prediction tasks. In Swin transformer methods, the input images are first split into nonoverlapping equal-sized rectangular patches by a patch splitting module like ViT, then they adopt a windowing configuration (W) that is shifted from that of the preceding layer and shifted the window (SW) to obtain a sequence of patches with different sizes. The process can be formulated as follows:

$$\begin{cases} \hat{z}^l = \text{WMSA}\left(\text{LN}\left(z^{l-1}\right)\right) + z^{l-1} \\ z^l = \text{MLP}\left(\text{LN}\left(\hat{z}^l\right)\right) + \hat{z}^l \\ \hat{z}^{l+1} = \text{SWMSA}\left(\text{LN}\left(z^l\right)\right) + z^l \\ z^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \end{cases} \tag{2}$$

where $\hat{z}^l$ and z^l denote the output features of the (S)WMSA module and the MLP module for block l, respectively.

Xie et al. [19] combined CNN and ViT for image segmentation. Zhou et al. [20] used interleaved convolution and self-attention operations in a transformer for volumetric medical image segmentation. Part of our work's inspiration results from the work of Yang et al. [12]. They applied the transformer to image super-resolution and proposed a texture transformer network for image super-resolution (TTSR). Chen et al. [21] proposed RegionViT. They replaced the self-attention in ViT with regional-to-local attention, which first extracts global information by regional self-attention and then exchanges the information among regional tokens with the associated local tokens via local self-attention. In this way, their method can inherit both global and local information, and thus outperform ViT in various computer vision tasks.

In summary, transformers have achieved success in various computer vision tasks. Nevertheless, most transformer methods split the input images into equal-sized rectangular patches, without consideration of the structural differences of different regions and semantic variations in the images. While the hierarchical PVT and Swin transformer methods overcome the limitation of equal-sized patches in vanilla transformers, they still use rectangular patches without considering the various shapes of different structures in an image. Compared with existing transformer methods, our DRT can learn to adaptively split the images into deformable structural regions with different sizes and shapes, which makes it suitable for the nonrigid registration of medical images that usually include multiple regions (organs) with structures of different sizes and shapes.

B. Deformable Related Methods

There have been many works attempting to split an image into deformable patches instead of rectangular patches [14], [22], [23], [24], [25], [26], [27], [28]. These methods often learn offsets to deform the patches. In the image classification task, Fu et al. [26] learned offsets of the most important regions in an image for better performance. Zhu et al. [27] proposed DCN to learn offsets for each pixel of the convolution kernel. Both of the above-mentioned methods use a convolution layer to produce the offsets. Deformable-DETR [28] applies deformable operations in the self-attention layers and cross-attention layers of DETR. However, it still relies on feature maps extracted from CNNs. DPT [14] is the work most close to ours, it proposed a DePatch module to learn offsets of patches and apply deformable patches in a transformer architecture. The DePatch module mainly consists of a linear layer. First, for a pixel $p(x, y)$ in a patch, the learned offset (δ_x, δ_y) and pixel coordinates (x', y') after deformation are as follows:

$$\left(\delta_x, \delta_y\right) = \tan(W_{\text{offset}} \cdot f(p)) \tag{3}$$

$$x' = x + \delta_x, \quad y' = y + \delta_y. \tag{4}$$

Here W_{offset} is the learnable weight of the offset, and $f(p)$ represents the function of the linear layer. The DPT method incorporates the DePatch module into the transformer and generates image features based on the deformed patches as follows:

$$z = W \cdot \text{Concat}\left(A\left(x', y'\right)\right) + b. \tag{5}$$

Here W and b are the learnable weights and bias of the network, respectively, while A and z represent the input and output features of the network, respectively.

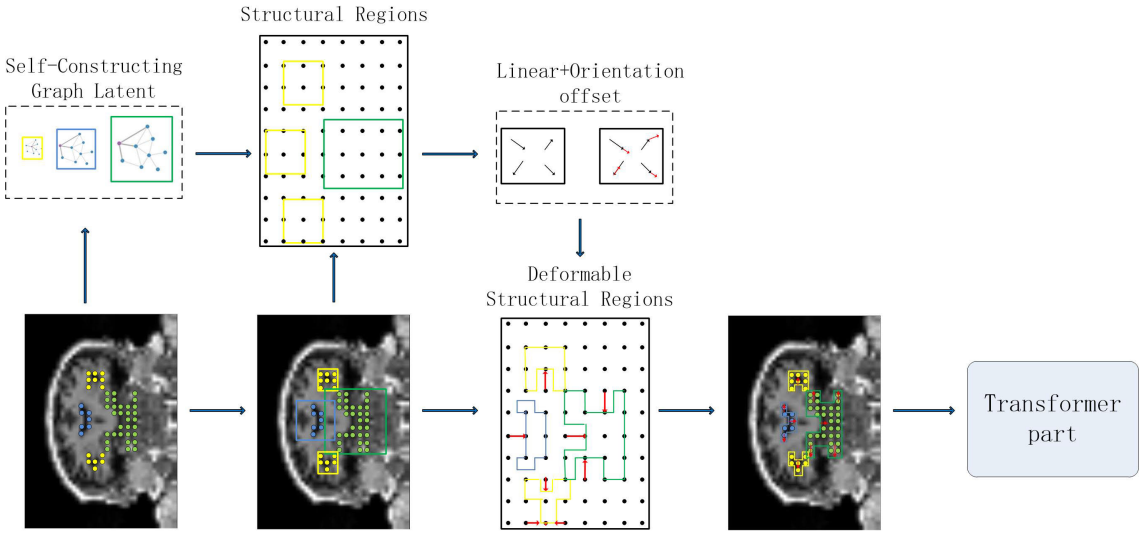


Fig. 2. Illustration of the DRSRE module. The module searches for the corresponding transformation relationship between the structural regions, by finding the most relevant ones and their sizes in images. 3-D images are simplified into two dimensions to represent an example of the embedding process.

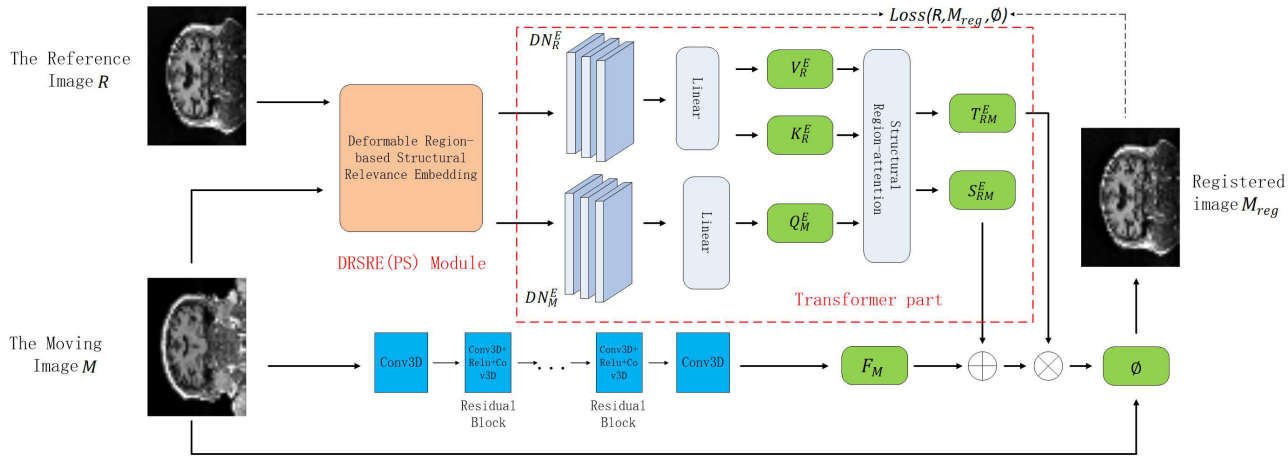


Fig. 3. Flowchart of the DRT for nonrigid medical image registration with a constraint of orientation. First, the input images R and M are embedded by the DRSRE module into deformable structure regions DN_R^E and DN_M^E. Then, in the transformer part, we propose to use linear layers and a structural region-attention module to calculate the structural region-attention map S_{RM}^E and a transferred feature map T_{RM}^E. Finally, several convolution blocks are used for abstracting features F_M of the moving image. We fuse F_M with T_{RM}^E and element-wisely multiply the fused feature by S_{RM}^E to obtain the deformation field ϕ.

In conclusion, the deformable patch-based networks replace the rectangular patch embedding module in vanilla transformers, so that it can learn to adaptively deform the patches to preserve the semantics of objects in the patches. However, since the offsets are learned by a convolution layer or linear layer based on equal-sized patches, they are still unable to overcome the limitation of equal-sized patches and cannot reflect structural differences in size of different regions in an image. In addition, they do not consider the structural directionality in patches. In our DRT method, we use an orientation constraint to learn the offsets for deformable structural regions and extract features better representing the semantic consistency of regions in medical images.

III. Method

In this section, we introduce the proposed DRT framework, which integrates the DRSRE module into a transformer part to obtain the accurate deformation field for nonrigid medical image registration. In the transformer part, we propose to use linear layers and a structural region-attention module to calculate the structural deformation relationship. Finally, a feature generator is used for abstracting features of the moving image. The abstracted features are corporate with the structural deformation relationship to obtain the deformation field and transform the moving image into the registered image. The overall flowchart of the proposed method is shown in Fig. 3.

A. DRSRE Module

Considering the shortcomings of the existing patch embedding methods, we propose a DRSRE module to preserve the semantic consistency in regions to obtain accurate transformation relationships for nonrigid image registration. The DRSRE module can adaptively split an image into deformable structural regions of different sizes and shapes. First, we adaptively

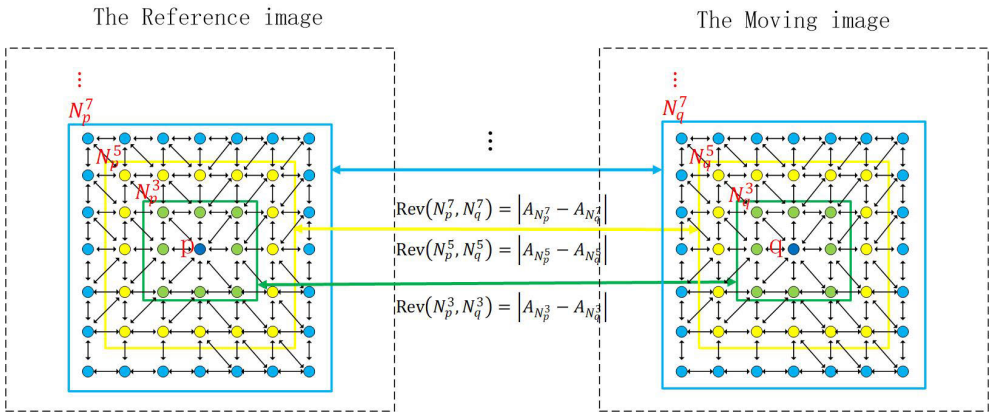


Fig. 4. Illustration of the deformable region-based structural relevance searching process. The module searches for the corresponding transformation relationship between the structural regions, by finding the most relevant ones and their sizes in images. 3-D images are simplified into two dimensions to represent an example of the embedding process.

search for the sizes of the deformable structural regions by the similarity of the self-constructing graph latent. As shown in Fig. 2, the search progress is based on the relevance of structural regions in the input images. Specifically, for a pixel p in the input reference image R and a pixel q in the moving image M, we extract a sequence of structural regions N_p and N_q with different sizes centered by the pixel as follows:

$$N_p = \left[N_p^3; N_p^5; \ldots; N_p^L\right], \quad p \in R \tag{6}$$

$$N_q = \left[N_q^3; N_q^5; \ldots; N_q^L\right], \quad q \in M. \tag{7}$$

Here L is the maximum size of the structural regions. Since the sizes of the input images for a registration task are the same, we assume that the size is $H \times W \times D$, and take the minimum value of the three dimensions as the maximum size of the structural regions as follows:

$$L = \min(W, H, D). \tag{8}$$

To obtain the similarity of the self-constructing graph latents in the structural regions, we compute adjacency matrices A_{N_p}, A_{N_q} of N_p, and N_q, respectively. The concept of the adjacency matrix comes from graph theory [29], which represents the structure of data by values of the edges between node points. Unlike image pixels, a graph can represent the image with nodes and edges, which is a better reflection of the geometry structure information. As a graph usually does not have a fixed grid, the nodes are not necessary to be spatially adjacent pixels, and thus have the ability to express more general spatial correspondence between images.

A graph represents an image in forms of nodes and edges, Specifically, for an image I, the graph G can be represented by

$$G = (N, E). \tag{9}$$

Here, N means the nodes, and E means edges of the graph. For convenience, the adjacency between nodes is usually represented by an adjacency matrix A. In this article, we follow the operations in the sklearn [30] library to obtain graphs of structural regions. First, we cluster the pixels of the structural region to obtain the key points and use these points as nodes. Then, we weigh the edges with the gradient values between the nodes. The adjacency matrix A_{N_p} and A_{N_q} of the structural regions are calculated as follows:

$$A_{N_{p(i,j)}} = \begin{cases} \mathrm{grad}\left(N_{pi}, N_{pj}\right), & \text{if } N_{pi} \text{ is adjacent to } N_{pj} \\ 0, & \text{if } N_{pi} \text{ is not adjacent to } N_{pj} \end{cases} \tag{10}$$

$$A_{N_{q(i,j)}} = \begin{cases} \mathrm{grad}\left(N_{qi}, N_{qj}\right), & \text{if } N_{qi} \text{ is adjacent to } N_{qj} \\ 0, & \text{if } N_{qi} \text{ is not adjacent to } N_{qj}. \end{cases} \tag{11}$$

Here N_{pi}, N_{pj}, N_{qi}, and N_{qj} represent pixel points in the ith and jth positions of structure regions N_p and N_q, grad$(\cdot)$ represents the gradient function, while $A_{N_{p(i,j)}}$ and $A_{N_{q(i,j)}}$ represent values in the (i, j) positions of adjacency matrices A_{N_p} and A_{N_q}, respectively.

Since the clustering results for each region are different, the structures of graph nodes and adjacency matrices of different regions are not always the same. In order to ensure that the number of graph nodes of the corresponding region in a pair of reference image and moving image stays the same, we apply a key point matching method [31] to filter the graph nodes and keep only the matched graph nodes in the corresponding regions. Compared with concatenation of vectors with fixed structure, the structure of adjacency matrix can be different according to the different clustering results in different regions, so graphs can better represent various structure information of different regions in the image.

In the next step, the relevance between the structural regions is evaluated via the distance between the adjacency matrix A_{N_p} and A_{N_q}, here we use the Frobenius norm [32] to measure the distance. Specifically, it is calculated by the square root of the sum of the differences between each element in A_{N_p} and A_{N_q} as follows:

$$\begin{aligned} \mathrm{Rev}\left(N_p, N_q\right) &= \left\| A_{N_p} - A_{N_q} \right\| \\ &= \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \left(A_{N_{p(i,j)}} - A_{N_{q(i,j)}}\right)^2 \right)^{1/2}. \end{aligned} \tag{12}$$

Here n is the number of elements in each dimension of the adjacency matrix.

As shown in (12), a smaller $\mathrm{Rev}(N_p, N_q)$ indicates higher relevance between the two regions, thus we obtain the most relevant structural regions N_p^E and N_q^E and their corresponding sizes E_p^q as follows:

$$E_p^q = \min\left(\mathrm{Rev}\left(N_p^e, N_q^e\right)\right), \quad e = 3, 5, \ldots, L. \tag{13}$$

Second, we predict an offset (δ_x, δ_y) with direction constraint which allows the structural regions to be deformable. For a pixel p in a structural region, we denote its coordinate as (x, y). The new coordinate of this pixel can be expressed as follows:

$$x' = x + \delta_x, \quad y' = y + \delta_y. \tag{14}$$

As shown in Fig. 4, we learn to predict the offset by adding a new branch, which consists of a linear layer and an orientation constraint. Based on the input structural regions, we predict $(\delta_{px}, \delta_{py})$ with (15) first, and then embed them with deformable structural regions DN_p and DN_q

$$\left(\delta_{px}, \delta_{py}\right) = \tan\left(W_{\text{offset}} \cdot \left(f(p) + \frac{p(x, y)}{\|\nabla_p\|} \cdot \theta_p\right)\right) \tag{15}$$

$$\mathrm{DN}_{p(x,y)} = N_{p\left(x+\delta_{px}, y+\delta_{py}\right)}, \quad \mathrm{DN}_{q(x,y)} = N_{q\left(x+\delta_{qx}, y+\delta_{qy}\right)}. \tag{16}$$

Here W_{offset} is the learnable weight of the offset, $f(p)$ represents a function of the linear layer, $p(x, y)$ is the pixel value at (x, y), while $\|\nabla_p\|$ and θ_p are the gradient value and orientation of pixel p, respectively. According to our previous work in [34], ∇_p and θ_p can be calculated by the following equations:

$$\begin{cases} \nabla_{p_x}(x, y) = p(x+1, y) - p(x, y) \\ \nabla_{p_y}(x, y) = p(x, y+1) - p(x, y) \end{cases} \tag{17}$$

$$\begin{cases} \theta_p = \arctan\left(\dfrac{\nabla_{p_x}^2 - \nabla_{p_y}^2}{2\,\nabla_{p_x}\,\nabla_{p_y}}\right)/2 \\ \theta_{p_x} = \cos\theta_p, \quad \theta_{p_y} = \sin\theta_p. \end{cases} \tag{18}$$

As we use the sliding window method to determine the central pixel of a deformable structural region, the process of the DRSRE module can be quite time-consuming. To solve that problem, a PS searching strategy is applied in the algorithm, which is described in Section III-E.

B. Structural Region-Attention

After the deformable structural regions are obtained via the DRSRE module, we propose a structural region-attention module in the transformer part to calculate the structural transformation relationship between the regions. Specifically, we transfer images based on the corresponding relationship of the deformable structural regions DN_R^E and DN_M^E, which are outputs of the DRSRE module as follows:

$$\mathrm{DN}_R^E = \left[\mathrm{DN}_p^E;\, p \in R\right] \tag{19}$$

$$\mathrm{DN}_M^E = \left[\mathrm{DN}_q^E;\, q \in M\right] \tag{20}$$

where p and q represent the central pixels of the deformable structural regions. We use linear layers to learn features for outputs of the DRSRE module, so the process can be expressed as follows:

$$\left(\left(K_R^E, V_R^E\right), Q_M^E\right) = \mathrm{Linear}\left(\mathrm{DN}_R^E, \mathrm{DN}_M^E\right) \tag{21}$$

where $\mathrm{Linear}(\cdot)$ denotes outputs of the linear layers, (K_R^E, V_R^E) are features for DN_R^E, and Q_M^E represents the feature for DN_M^E. To obtain the deformation relationship between DN_R^E and DN_M^E, we propose a structural region-attention module and use Q_M^E, K_R^E, V_R^E as inputs of the module. For each feature Q_q^E in Q_M^E and K_p^E in K_R^E, we first calculate the relevance $r_{p,q}^E$ between the input features by the normalized inner product as follows:

$$r_{p,q}^E = \left\langle \frac{Q_q^E}{\left\|Q_q^E\right\|}, \frac{K_p^E}{\left\|K_p^E\right\|} \right\rangle, \quad p \in \mathrm{DN}_R^E,\ q \in \mathrm{DN}_M^E. \tag{22}$$

Then, a structural region-attention map R_{pq} is computed from $r_{p,q}^E$ to represent the most relevant position in the features

$$R_{pq} = \arg\max_q r_{p,q}^E, \quad p \in \mathrm{DN}_R^E,\ q \in \mathrm{DN}_M^E. \tag{23}$$

Inspired by the TTSR [33] network, we compute the transferred feature map T_{RM}^E and the confidence map S_{RM}^E of the transferred features as follows:

$$T_{\mathrm{RM}}^E = V_{R_{pq}}^E, \quad p \in \mathrm{DN}_R^E,\ q \in \mathrm{DN}_M^E \tag{24}$$

$$S_{\mathrm{RM}}^E = \max_q r_{p,q}^E, \quad p \in \mathrm{DN}_R^E,\ q \in \mathrm{DN}_M^E. \tag{25}$$

As a result, we obtain a transferred feature map T_{RM}^E and its confidence map S_{RM}^E as outputs of the structural region-attention module.

C. Feature Generation and Transformation

Instead of directly applying S_{RM}^E and T_{RM}^E to obtain the deformation field, we first use a feature generator to extract features F_M from the moving image and then fuse the extracted features with the transferred feature map T_{RM}^E to leverage more local information from the moving image. The feature generator consists of several convolution layers and residual blocks. The architecture of the module is shown in Fig. 5.

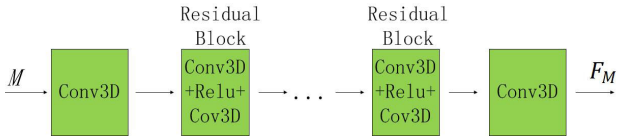


Fig. 5. Feature generation module.

An input image M is first passed through a 3-D convolution layer and is then processed by several residual blocks. A residual block includes two 3-D convolution layers and a ReLU layer. Finally, the output of the residual blocks goes through a 3-D convolution layer, and the extracted feature vector F_M is obtained. The parameters of the module are shown in Table I.

With the fused feature map of F_M and T_{RM}^E, we element-wisely multiply it by the confidence map S_{RM}^E and obtain the deformation field ϕ, the process can be presented as follows:

$$\phi = \mathrm{Conv}\left(\mathrm{Concat}\left(F_M, T_{\mathrm{RM}}^E\right)\right) \cdot S_{\mathrm{RM}}^E \tag{26}$$

$$M_{\mathrm{reg}} = M \odot \phi. \tag{27}$$

TABLE I
ARCHITECTURE OF THE FEATURE GENERATION MODULE. Conv(N_{in}, N_{out}) REPRESENTS THE 3-D CONVOLUTIONAL LAYER WITH N_{in} INPUT CHANNELS AND N_{out} OUTPUT CHANNELS. THE KERNEL SIZE IS $3 \times 3 \times 3$ FOR ALL CONVOLUTIONAL LAYERS

Id	Layer Name	
0	Conv(3,64)	
1	Residual Block	Conv(64,64)
		ReLU
		Conv(64,64)
2	Residual Block	Conv(64,64)
		ReLU
		Conv(64,64)
...	Residual Block	Conv(64,64)
		ReLU
		Conv(64,64)
64	Residual Block	Conv(64,64)
		ReLU
		Conv(64,64)
65	Conv(64,3)	

Here the operator $\cdot$ denotes elementwise multiplication, Conv and Concat represent a convolutional layer and the concatenation operation, respectively, $\odot$ represents the warping operation, and M_{reg} is the final output of the DRT algorithm.

D. Loss Functions

As the aim of nonrigid registration is to warp the moving image to align with the reference image, it can be regarded as an optimization problem written as

$$\hat{\phi} = \arg\min_{\phi} L\big(R, M_{\text{reg}}, \phi\big). \tag{28}$$

Here R is the reference image, M is the moving image, ϕ is the deformation field, M_{reg} is M warped by ϕ, and $L(\cdot)$ is the objective function. As proven by Vercauteren et al. [35], the objective function $L(\cdot)$ of the nonrigid registration should include the similarity measure and the smoothness restriction of the deformation field

$$L\big(R, M_{\text{reg}}, \phi\big) = L_{\text{sim}}\big(R, M_{\text{reg}}\big) + L_{\text{smooth}}(\phi). \tag{29}$$

$L_{\text{sim}}(\cdot)$ measures similarity between the warped moving image M_{reg} and the reference image R. $L_{\text{smooth}}(\cdot)$ imposes smoothness regularization on the deformation field ϕ.

There are several common formulations for ϕ, L_{sim} and L_{smooth}. Often, the deformation field ϕ is a displacement vector field, specifying the vector offset from M to R for each voxel. $L_{\text{sim}}(R, M_{\text{reg}})$ enforces the warped moving image M_{reg} to be similar with the reference image R, so as to obtain the accurate deformation field ϕ. Common metrics used for L_{sim} include mean squared difference, mutual information, and cross correlation. The latter two are useful when images have varying modes and intensity contrasts. Since we focus on nonrigid registration of medical images of the same modality (MRI) in this work, the mean squared difference between R and M_{reg} is used as the L_{sim} constraints in the loss function

$$L_{\text{sim}}\big(R, M_{\text{reg}}\big) = \frac{1}{\Omega}\sum_{p\in\Omega}\big|R(p) - M_{\text{reg}}(p)\big|^2. \tag{30}$$

Here p denotes the pixel location, and Ω represents the image domain. $L_{\text{smooth}}(\phi)$ constrains the smoothness of deformation field ϕ to preserve topology in images, which is often modeled as a linear operator on gradients of ϕ. In this work, we present the L_{smooth} in loss function as

$$L_{\text{smooth}}(\phi) = \|\phi\|. \tag{31}$$

$\|\cdot\|$ is the spatial gradient operation. The total loss function for DRT network training can be presented as

$$L\big(R, M_{\text{reg}}, \phi\big) = L_{\text{sim}}\big(R, M_{\text{reg}}\big) + L_{\text{smooth}}(\phi). \tag{32}$$

E. Progressive Searching

In the DRSRE module, a sliding window method is used to determine the central pixel of a structural region. However, this process is quite time-consuming and requires excessive memory resources. To improve the efficiency of the proposed algorithm, we explore a PS strategy that adaptively reduces the range of the searching area during the process. The main flow of the strategy is shown in Fig. 6.

The medical images used for deformable registration in this article are 3-D. For convenience, we only use two dimensions to represent an example of the PS process. In the direction of the horizon, first, for an initial center pixel p_{11}, we obtain the most relevant size E_{11} by the DRSRE module and represent the pixel as $p_{11}^{E_{11}}$. Then, instead of searching for the next structural region from the neighborhood pixel, we set the next center pixel p_{12} as pixel $2 \times E_{11}$ away from pixel p_{11}. After obtaining the most relevant size E_{12} of pixel p_{12} by the DRSRE module, pixel $2 \times E_{12}$ away from pixel p_{12} is set as the next center pixel. It can be seen that the central pixel of a structural region is progressive searched according to the size of the previous region and the size of each region is determined adaptively according to the size of the semantic structure around the central pixel, so the structural regions in an image can adaptively cover all of the structures without missing effective local information. When there are background regions between the semantic structures, there will exist uncovered regions between the structural regions. In medical images, since the pixel values of these background regions are usually zeros and the noise in these regions may interfere with the registration result, so we just discard these ineffective background regions in the searching progress. A similar strategy is applied in other dimensions. Thus, unnecessary calculation of the overlapping parts between different structural regions can be avoided. In addition, when the size of the structural regions exceeds the image boundary, the structural regions will become incomplete. In order to ensure the integrity of each region, we pad these regions with zero so as to facilitate the subsequent calculation. From another perspective, the efficiency of the algorithm is dramatically improved because of the reduction of the computation.

IV. EXPERIMENTS

We evaluate the effectiveness of DRT for nonrigid medical image registration and compare its performance to various state-of-the-art baseline methods. Our goal is for DRT to

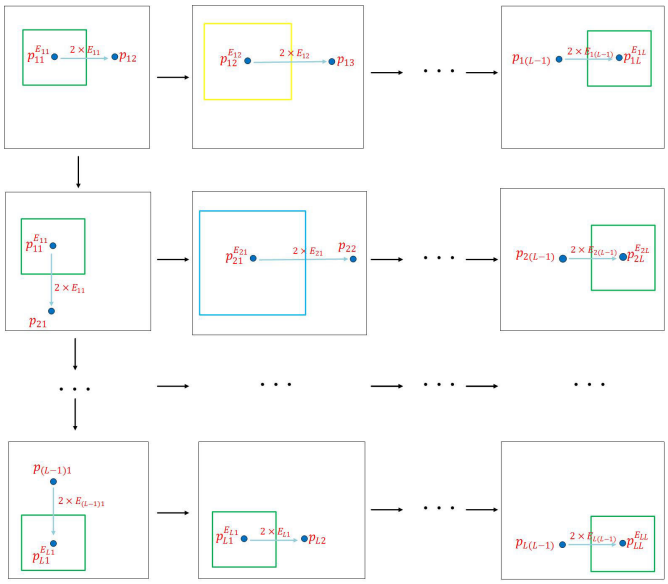


Fig. 6. Illustration of the progressive searching process. 3-D images are simplified into two dimensions to represent an example of the search process.

obtain an accurate deformation field of the input moving images, matching or exceeding the accuracy of all nonlearning and learning baseline methods. The structure of this section is as follows. In Section IV-A, details of the two databases are introduced. Section IV-B illustrates the baseline methods. Section IV-C shows implementation details for the two datasets. Section IV-D introduces the evaluation metrics. The experimental results and ablation study are presented in Sections IV-E and IV-G, respectively. Finally, we discuss the semantic inconsistency and computational efficiency in Sections IV-F and IV-H.

A. Datasets

We evaluate the DRT-based registration method on two 3-D medical image datasets, named the LPBA40 [36] database and ADNI [37] database. Both datasets contain many medical images that have multiple regions of interest (ROIs) with different structural information.

The LPBA40 database consists of 3-D T1 weighted MRI images from 40 young adults, and each image has 54 ROI sections. Each image is normalized via affine transformation [38] and converted to a common MNI152 space [39]. The image intensity is normalized by histogram matching [40]. The LPBA40 database is mainly used to study nonrigid registration of ROI sections between different images. To improve computational efficiency, each subject's 3-D image is resized to 25% of the original size as $64 \times 64 \times 64$ before being fed into the network.

The ADNI database includes 199 3-D T1 weighted brain MRI scans. The dataset contains normal and lesion brain images of Alzheimer's disease from 29 subjects. Each subject has a normal image and a series of corresponding lesion images are acquired at different times. Each scan is extracted and normalized using FreeSurfer [41]. The purpose of the ADNI database is to study changes in brain structures by comparing deformation in different pathological stages. Similar to the LPBA40 database, we resize the images to 25% of the original size as $40 \times 48 \times 56$ before feeding them into the network.

B. Baseline Methods

We compare the proposed DRT-based registration method with fifteen state-of-the-art baseline methods. For nonlearning-based methods, four representative classical nonrigid registration methods are used as baselines. We test the SyN [42] using recommended parameters and mean squared difference as the objective function, along with a default Gaussian smoothing of 3. We test NiftyReg [43] with the default sum of squared differences as the objective function and a regularization weighting of 0.0002. We test the deedsBCV [44] by using self-similarity context as the objective function. To make it suitable for the spatial scales of brain structures, we reduce the default grid spacing, search radius, and quantization step to $8 \times 7 \times 6 \times 5 \times 4$, $8 \times 7 \times 6 \times 5 \times 4$, and $5 \times 4 \times 3 \times 2 \times 1$, respectively. We also test the LDDMM [45] using the mean squared error as the objective function with a smoothing kernel size of 2.

For state-of-the-art learning-based nonrigid registration methods, four CNN-based methods, six transformer-based methods, and one deformable-patch-based method are used as baselines. To compare the proposed method with CNN-based methods, we train VoxelMorph [46] using normalized cross correlation as the loss function and set the regularization parameter to 0.02. The trained model is then tested on pairs of images from the test sets of the LPBA40 database and ADNI database. We train and test VoxelMorph-diff [48] with the same loss function as VoxelMorph. The regularization parameter is set to 0.01. We train and test CycleMorph [49] by setting the weights for cycle loss, identity loss, and deformation field regularization to 0.1, 0.5, and 0.02, respectively. We train and test MIDIR [50] with the same loss function and

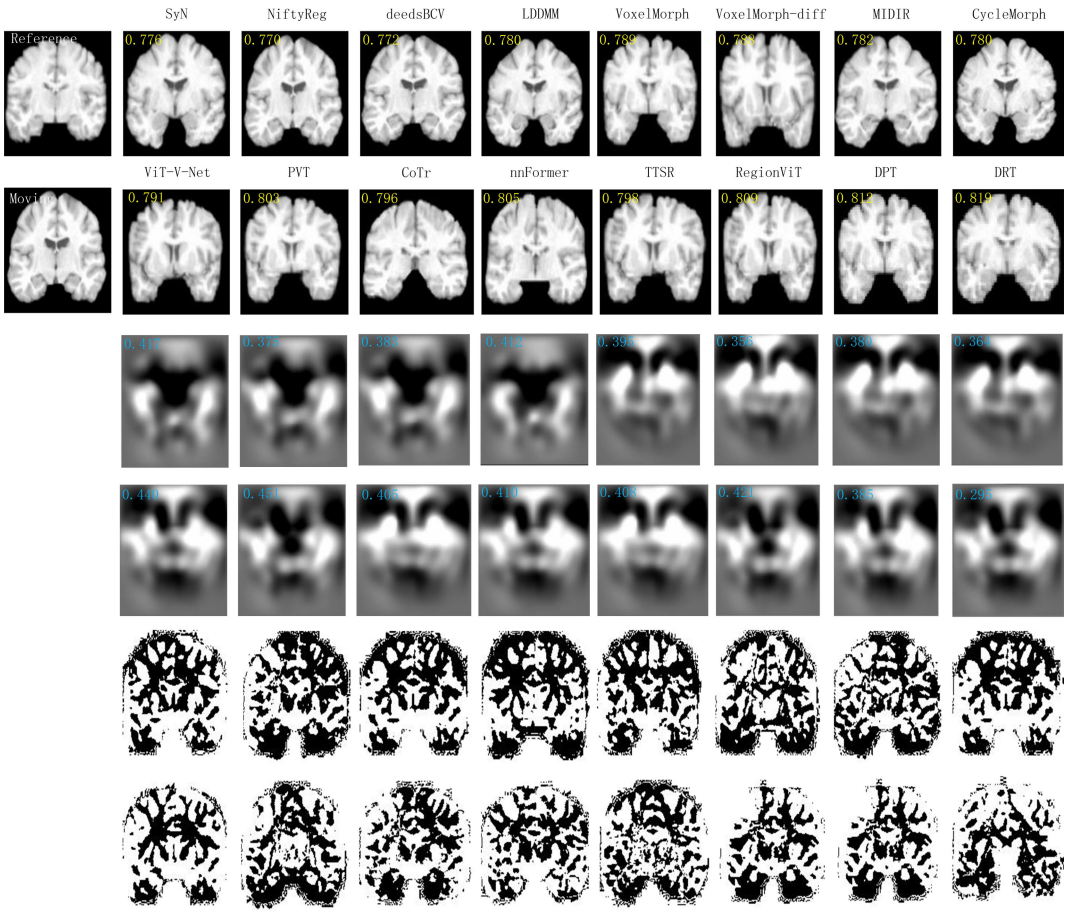


Fig. 7. Qualitative results of a sample slice for nonrigid image registration on the test set of LPBA40. The top two rows show the registered images, and the dice scores are averaged for all the ROI sections and marked in yellow. The middle two rows exhibit the corresponding deformation fields, and percentages of $|J_\phi| \leq 0$ are marked in blue. The bottom two rows show the differences map between the registered image and the moving image.

TABLE II

QUANTITATIVE EVALUATION RESULTS OF NONRIGID REGISTRATION BASED ON THE TEST SET OF LPBA40. DICE SCORES AND PERCENTAGES OF $|J_\phi| \leq 0$ ARE EVALUATED FOR DIFFERENT METHODS. BOTH OF THE EVALUATION METRICS ARE AVERAGED FOR ALL THE ROI SECTIONS IN IMAGES, AND THE BEST ONES ARE SHOWN IN BOLD

Method	Dice	$\%of\ \lvert J_\phi\rvert \leq 0$
No registration	0.623 ± 0.224	-
SyN[42]	0.710 ± 0.054	0.618 ± 0.021
NiftyReg[43]	0.701 ± 0.102	0.489 ± 0.024
deedsBCV[44]	0.697 ± 0.049	0.511 ± 0.034
LDDMM[45]	0.706 ± 0.053	0.614 ± 0.018
VoxelMorph[46]	0.718 ± 0.037	0.574 ± 0.019
VoxelMorph-diff[48]	0.710 ± 0.056	0.328 ± 0.012
CycleMorph[49]	0.719 ± 0.047	0.441 ± 0.011
MIDIR[50]	0.716 ± 0.032	0.500 ± 0.039
ViT-V-Net[15]	0.720 ± 0.149	0.775 ± 0.010
PVT[17]	0.715 ± 0.217	0.894 ± 0.011
CoTr[19]	0.716 ± 0.035	0.611 ± 0.018
nnFormer[20]	0.718 ± 0.319	0.618 ± 0.019
TTSR[12]	0.719 ± 0.318	0.615 ± 0.010
RegionViT[21]	0.723 ± 0.033	0.626 ± 0.011
DPT[14]	0.730 ± 0.033	0.510 ± 0.037
DRT	$\mathbf{0.748 \pm 0.020}$	$\mathbf{0.217 \pm 0.016}$

regularization parameter as VoxelMorph. All settings of the parameters follow the default recommended parameters.

To evaluate the proposed DRT-based registration method, we compare it with state-of-the-art transformer-based methods in computer vision. We modify these methods to produce a 3-D deformation field that warps the input moving image to obtain the registered image. While a spatial transformation and deformation layer is added to these methods, the parameters

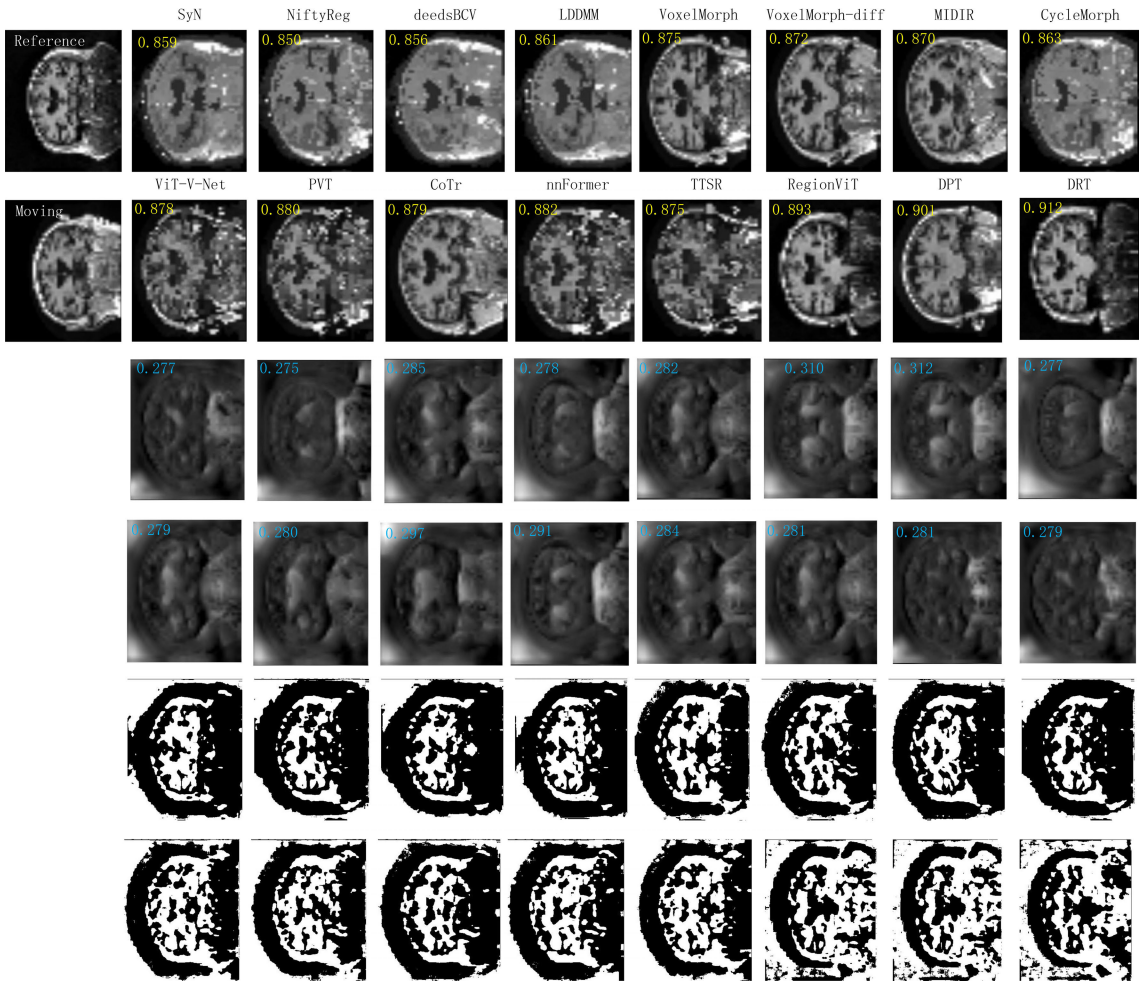


Fig. 8. Qualitative results of a sample slice for nonrigid image registration on the test set of ADNI. The top two rows show the registered images, and the dice scores are averaged for all the ROI sections and marked in yellow. The middle two rows exhibit the corresponding deformation fields and percentages of $|J_\phi| \leq 0$ are marked in blue. The bottom two rows show the differences map between the registered image and the moving image.

TABLE III

QUANTITATIVE EVALUATION RESULTS OF NONRIGID REGISTRATION BASED ON THE TEST SET OF ADNI. DICE SCORES AND PERCENTAGES OF $|J_\phi| \leq 0$ ARE EVALUATED FOR DIFFERENT METHODS. BOTH OF THE EVALUATION METRICS ARE AVERAGED FOR ALL THE ROI SECTIONS IN IMAGES, AND THE BEST ONES ARE SHOWN IN BOLD

Method	Dice	$\%of\ \|J_\phi\| \leq 0$
No registration	0.809 ± 0.088	-
SyN[42]	0.843 ± 0.130	0.851 ± 0.109
NiftyReg[43]	0.839 ± 0.204	0.551 ± 0.089
deedsBCV[44]	0.841 ± 0.108	0.310 ± 0.024
LDDMM[45]	0.840 ± 0.114	0.433 ± 0.058
VoxelMorph[46]	0.851 ± 0.109	0.501 ± 0.069
VoxelMorph-diff[48]	0.848 ± 0.199	0.622 ± 0.056
CycleMorph[49]	0.846 ± 0.294	0.635 ± 0.067
MIDIR[50]	0.849 ± 0.135	0.517 ± 0.054
ViT-V-Net[15]	0.855 ± 0.120	0.511 ± 0.040
PVT[17]	0.859 ± 0.097	0.514 ± 0.048
CoTr[19]	0.864 ± 0.113	0.575 ± 0.044
nnFormer[20]	0.857 ± 0.110	0.486 ± 0.050
TTSR[12]	0.856 ± 0.214	0.459 ± 0.062
RegionViT[21]	0.867 ± 0.121	0.484 ± 0.034
DPT[14]	0.870 ± 0.133	0.411 ± 0.032
DRT	$\mathbf{0.886 \pm 0.073}$	$\mathbf{0.308 \pm 0.030}$

and most of the network architectures remain unchanged. To make the number of patches close to the number of ROIs in the original images, the sizes of patches are set to 3. We train and test ViT-V-Net [15], PVT [17], CoTr [19], nnFormer [20] and RegionViT [21] with the default parameters. For TTSR [12], we set the moving image as the LR image and the

reference image as the reference image before obtaining the output of the transformer as the registered image. For DPT [14], we incorporate the DePatch module into ViT and use it for nonrigid image registration.

C. Implementation Details

The proposed DRT is implemented by PyTorch [51] using a batch size of 1 with 12G NVIDIA TITAN GPU. All models are trained with the Adam optimizer [52] with a learning rate of 10^{-4}. The transformer parameters of DRT follow the default parameters of ViT-V-Net [15].

The networks are trained until they converge and are tested in a pairwise fashion, which allows the moving image and the reference image to be arbitrary. The dataset is split into training, validation, and testing sets according to the proportion of 7:1:2, which follows the setting of ViT-V-Net [15]. Therefore, there are $28 \times 27/4 \times 3/8 \times 7$ registrations for the LPBA40 dataset and $139 \times 138/20 \times 19/40 \times 39$ registrations for the ADNI dataset in total.

D. Evaluation Metrics

The registration performance of the models is evaluated based on the overlap between anatomical segmentation, which can be represented by computing the Dice score [53] across the ROIs. Given the registered moving image M_{reg} and the reference image R, the Dice score of the two volumes is calculated as follows:

$$\text{Dice}\left(M_{\text{reg}}, R\right) = 2 * \frac{\left|M_{\text{reg}} \cap R\right|}{\left|M_{\text{reg}}\right| + |R|}. \tag{33}$$

The Dice score ranges between 0 and 1, which corresponds to the coincidence of two images and thus reflects the quality of registration. A Dice score closer to 1 shows that the registered result is relatively more identical, and a score of 0 represents that no overlap is obtained.

We averaged the Dice scores of all the ROI regions to obtain an average score. The mean and standard deviation of the averaged scores of all images are compared across various registration methods.

In the registration process, especially between images of the same patient, since the semantics and topology of images are not supposed to be changed, image folding should not occur anatomically. We can verify whether there are foldings in the registration by quantitatively evaluating the smoothness of the deformation field, or by visualizing the Jacobian matrix, so as to judge the quality of the registration results. According to the definition of Jacobian, the percentages of nonpositive values in the determinant of the Jacobian matrix for the deformation fields (i.e., $|J_\phi| \leq 0$) [54] can be computed as follows:

$$\left|J_{\phi(i,j,k)}\right| = \begin{vmatrix} \frac{\delta i}{\delta x} & \frac{\delta j}{\delta x} & \frac{\delta k}{\delta x} \\ \frac{\delta i}{\delta y} & \frac{\delta j}{\delta y} & \frac{\delta k}{\delta y} \\ \frac{\delta i}{\delta z} & \frac{\delta j}{\delta z} & \frac{\delta k}{\delta z} \end{vmatrix}. \tag{34}$$

Here i, j, k represent the ith, jth, and kth positions of a pixel point in the image, and x, y, z represent the spatial decomposition directions of the deformation field. When $|J_\phi| = 1$, the volume at this pixel does not change; $|J_\phi| > 1$, expansion occurs; $0 < |J_\phi| < 1$, shrinkage occurs; $|J_\phi| \leq 0$, image folding occurs. We can quantify the smoothness of the deformation field by calculating the percentages of $|J_\phi| \leq 0$ and the standard deviation of Jacobian in the registered images. The smaller the percentage of $|J_\phi| \leq 0$, the better the smoothness of the deformation field, and the higher the effectiveness of the registration method.

E. Results and Discussion

Table II lists the quantitative evaluation results of the nonrigid image registration experiments based on the test set of the LPBA40 dataset. The proposed DRT method produces the highest Dice score of 0.748 with the smallest foldings (i.e., 0.217% of $|J_\phi| \leq 0$) in the deformation fields. In comparison, DRT improved the Dice score by 4.2% when compared to the best CNN-based method (i.e., VoxelMorph) and by 5.4% when compared to the best nonlearning-based method (i.e., SyN). The transformer-based methods (i.e., ViT-V-Net, PVT, CoTr, nnFormer, TTSR, RegionViT, DPT) generally perform better than the CNN-based methods by overcoming the limitation of the convolution operation. Note that our DRT performs better than other transformer-based methods with a higher Dice score, demonstrating the DRSRE module's superiority over the equal-sized or rectangular patch embedding. Note that the Dice score of our DRT method is close to DPT while percentages of $|J_\phi| \leq 0$ of DRT is significantly smaller than that of DPT. The reason is that although DPT overcomes the limitations of rectangular patches, it still learns the offset based on equal-sized split patches and is therefore unable to reflect structural differences in the size of different regions in an image. In addition, it learns the offset by a single convolution layer, which does not consider the structural directionality in the regions. Compared with DPT, DRSRE overcomes the limitations of both the equal-sized and rectangular patch split, which can learn to adaptively split an image into deformable structural regions with different sizes and shapes according to the structural features with orientation constraints. DRT can represent the structural differences and directionality between regions and effectively preserve the semantic consistency in regions to obtain accurate transformation relationships for nonrigid image registration. Since Dice score focuses on the overall similarity of images and the Jacobian matrix focuses on the local consistency of structures in images, the percentage of image foldings ($|J_\phi| \leq 0$) for defomation field can well reflect the advantage of the DRSRE module and DRT-based registration method. The qualitative results of a sample slice for nonrigid image registration are shown in Fig. 7. The images are visualized via MRIcron, which is available at http://www.sph.sc.edu/comd/rorden/mricron. Among all the registration methods, DRT shows the best qualitative result with the smallest differences between the registered image and the moving image. On the other hand, DRT and RegionViT produce the smoothest deformations inside the brain area.

To further evaluate the DRT method, we conduct nonrigid image registration experiments on the ADNI dataset. The quantitative evaluation results are listed in Table III. The proposed DRT method achieves the highest Dice score of 0.886 with the smallest foldings (i.e., 0.308% of $|J_\phi| \leq 0$) in the deformation fields. Additionally, the transformer-based methods perform better than the CNN-based methods (i.e., VoxelMorph, VoxelMorph-diff, CycleMorph, and MIDIR) and the nonlearning-based methods (i.e., SyN, NiftyReg, deeds-BCV, and LDDMM). Since the proposed DRSRE module in our method adaptively embed an image into deformable structural regions of different sizes and shapes rather than using equal-sized or rectangular patches, the DRT-based registration method can well preserve the semantic and structural consistency across the structures in the image (see Fig. 1), so the smoothness of the deformation field is better and percentage of image foldings ($|J_\phi| \leq 0$) in the registration results of DRT is significantly lower than other state-of-the-art CNN-based and transformer-based nonrigid registration methods. Compared with Dice score index, which focuses more on the overall similarity of images, the Jacobian matrix focuses more on the local consistency of structures in images that can better reflect the advantage of the DRSRE module and DRT-based registration method. Fig. 8 shows the qualitative results of a sample slice for nonrigid image registration. Among all the methods, the DRT method produced the best qualitative results, with the smallest differences between the registered image and the moving image.

F. Semantic Consistency

In comparison with equal-sized or rectangular patch embedding in state-of-the-art methods (i.e., CNN-based VoxelMorph, transformer-based ViT-V-Net, and deformable-patch-based DPT), the proposed DRT use a DRSRE module to split the images into deformable structural regions of different sizes and shapes. In this way, DRT can effectively preserve the semantic consistency across the regions to calculate accurate transformation relationships for nonrigid image registration. To verify the correlation between semantic consistency and accuracy of nonrigid medical image registration, we compare semantic consistency in the registration results of DRT with VoxelMorph, ViT-V-Net, and DPT. Here, we use the edge continuity index LX [56] to quantify the semantic consistency across the regions in an image. Continuity is an important natural attribute of regional edges. Edge continuity determines whether the semantics of the regional structure is inconsistent. Fragmented and broken edges indicate that the integrity of the regional structure is poor. In many region-based image processing applications, such as image segmentation and object recognition, the continuity of region edges plays a decisive role. Therefore, as an important attribute of regions, edge continuity can be used as an index to evaluate structural semantic consistency. The index can be calculated as follows:

$$\mathrm{LX} = \frac{1}{\max(L_i X_i)} \sum_{i=1}^{n} L_i X_i / n. \tag{35}$$

Here n is the number of regions, L represents the length of the region edge, X denotes the convex hull area of the region edge, and $\max(L_i X_i)$ is the maximum of $L_i X_i$. The larger the index is the stronger the semantic inconsistency. Fig. 9 shows the average indices of each ROI in the results of the nonrigid image registration experiments based on the test set of the LPBA40 dataset, and similar findings could be found in the ADNI dataset. DRT obtains larger edge continuity indices in most of the ROIs than the other methods. While the indices of DPT are larger than those of VoxelMorph and ViT-V-Net, they are significantly smaller than the indices of DRT. The quantitative evaluation results are listed in Table IV. The index of DRT is 17.2% higher than that of DPT, 18.2% higher than that of ViT-V-Net, and 19.3% higher than that of VoxelMorph. It can be concluded that the semantic consistency in the registration results is a positive correlation with the accuracy of the nonrigid image registration, which verifies the superiority of the DRT method over other methods in preserving semantic consistency and its importance in nonrigid image registration.

G. Ablation Study

In this section, we present experiments designed to verify the effectiveness of the modules in the DRT architecture. Specifically, we carry out four ablation studies of the network on the test set of the LPBA40 dataset, and similar findings can be found in the ADNI dataset.

Table V lists the average results of the Dice scores in the LPBA40 dataset obtained by using: 1) DRT (DRSRE$^-$) where we use the DRT framework without the DRSRE module; 2) DRT (DRSRE*) where we use the DRT framework without the orientation constraint in the DRSRE module; 3) DRT (SRA$^-$) where we use the DRT framework without the structural region-attention module; and 4) DRT (FG$^-$) where we use the DRT framework without the feature generator of the moving image.

1) Analysis of the DRSRE: In the DRT (DRSRE$^-$) method, we replace the DRSRE module by directly using the input images for transformation. Specifically, for the input reference image R and the moving image M, we replace the output of the DRSRE module, DN_R^E, and DN_M^E, with the images R and M. Thus, (24)–(26) in DRT are written as follows:

$$T_{\mathrm{RM}} = V_R \tag{36}$$

$$S_{\mathrm{RM}} = \max_M \left\langle \frac{Q_M}{\|Q_M\|}, \frac{K_R}{\|K_R\|} \right\rangle \tag{37}$$

$$\phi = \mathrm{Conv}(\mathrm{Concat}(F_M, T_{\mathrm{RM}})) \cdot S_{\mathrm{RM}}. \tag{38}$$

To verify the effectiveness of the orientation constraint in the DRSRE module, we remove it in the DRT (DRSRE*) method, so (15) and (16) in DRT become as follows:

$$\left(\delta'_{px}, \delta'_{py}\right) = \tan(W_{\mathrm{offset}} \cdot f(p)) \tag{39}$$

$$\mathrm{DN}'_{p(x,y)} = N_{p\left(x+\delta'_{px}, y+\delta'_{py}\right)}, \quad \mathrm{DN}'_{q(x,y)} = N_{q\left(x+\delta'_{qx}, y+\delta'_{qy}\right)}. \tag{40}$$

Here δ'_{px} and δ'_{py} are the offsets predicted without the orientation constraint, while $\mathrm{DN}'_{p(x,y)}$ and $\mathrm{DN}'_{q(x,y)}$ are the corresponding deformable structural regions. In this way, (21)

TABLE IV

AVERAGE EDGE CONTINUITY INDICES OF ROIS IN NONRIGID REGISTRATION RESULTS BASED ON THE TEST SET OF LPBA40. INDICES OF ALL ROIS ARE AVERAGED INTO ONE INDEX FOR QUANTITATIVE EVALUATION

	VoxelMorph	ViT-V-Net	DPT	DRT
Indices	0.627	0.633	0.638	0.748

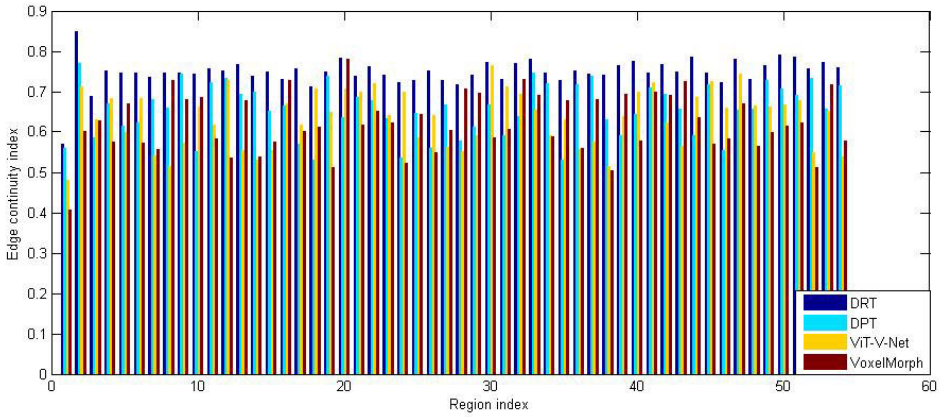


Fig. 9. Boxplot of edge continuity indices of 54 brain ROIs in the LPBA40 test dataset. Indexes of all images are averaged into one index for visualization.

TABLE V

ABLATION STUDY RESULTS BASED ON THE TEST SET OF LPBA40. THE DICE SCORES AND PERCENTAGE VALUES OF THE NONPOSITIVE JACOBIAN DETERMINANT ON THE DEFORMATION FIELDS ARE AVERAGED FOR ALL THE SUBJECTS

Method	Dice	$\%of\ \lvert J_\phi\rvert \leq 0$
No registration	0.623 ± 0.224	-
DRT($DRSRE^-$)	0.661 ± 0.023	0.206 ± 0.041
DRT($DRSRE^*$)	0.738 ± 0.021	0.214 ± 0.035
DRT(SRA^-)	0.720 ± 0.025	0.220 ± 0.028
DRT(FG^-)	0.741 ± 0.024	0.208 ± 0.021
DRT	0.748 ± 0.020	0.217 ± 0.016

TABLE VI

COMPUTATIONAL TIMES OF DIFFERENT IMAGE REGISTRATION ALGORITHMS

	DRT^*	SyN	VoxelMorph	ViT-V-Net	DPT	DRT
Time(s)	8.741	0.813	0.703	0.761	0.673	0.796

* DRT^* refers to the DRT framework without applying the PS strategy.

in DRT are written as follows:

$$\left(\left(K_R^E, V_R^E\right), Q_M^E\right) = \text{Linear}\left(\left(\text{DN}_p^E\right)', \left(\text{DN}_q^E\right)'\right), \quad p \in R, \quad q \in M. \tag{41}$$

The quantitative and qualitative experimental results of deformable registration in the former sections have shown the importance of the DRSRE module. This novel kind of deformable structural embedding proposed in this work overcomes the shortcomings of the current equal-sized or rectangular patch splitting methods by learning to adaptively split the input images into deformable structural regions with different sizes and shapes. According to Table V, it is clear that DRT (DRSRE$^-$) cannot achieve satisfactory registration results, with a lower Dice score and higher foldings in the deformation fields observed in comparison with the DRT-based registration method. DRT (DRSRE*) obtains a significantly higher Dice score but also obtains a high folding in the deformation fields with 0.224% of $|J_\phi| \leq 0$. This indicates that the orientation constraint plays an important role in smoothing the deformation fields. In conclusion, the DRSRE module enables the network to achieve great generalization capacity and obtains increased robustness, which allows the proposed method to outperform the other methods.

2) Analysis of the Structural Region-Attention: In the DRT (SRA$^-$) method, we remove the structural region-attention module and use the deformable structural regions directly for transformation. Specifically, for the input reference image R and the moving image M, we remove the output of the structural region-attention module, concatenate F_M and DN_R^E, and add the result back to F_M as the registered image M_{reg}. In this way, (26) in DRT is written as follows:

$$\phi = \text{Conv}\left(\text{Concat}\left(F_M, \text{DN}_R^E\right)\right). \tag{42}$$

As shown in Table V, we can find that DRT (SRA$^-$) performs worse than the DRT-based registration method. It is concluded that the structural region-attention is a vital module in the proposed method, and thus the images cannot be accurately registered without it. This part is a key step to obtain the corresponding transformation relationship between the deformable structural regions, and it is a necessary module for achieving satisfactory registration results.

3) Analysis of the Feature Generation: In the DRT (FG$^-$) method, we replace the feature generator of the moving image by directly using the input moving images for transformation.

Specifically, for the input reference image R and the moving image M, we replace the output of the feature generator, F_M, with the image M. In this way, (26) in DRT is written as follows:

$$\phi = \text{Conv}\left(\text{Concat}\left(M, T^E_{\text{RM}}\right)\right) \cdot S^E_{\text{RM}}. \tag{43}$$

In the fifth row of Table V, we can see that DRT (FG^-) performs slightly worse than the proposed network, but it functions better than DRT (DRSRE^-). This indicates that while the structural transformer has a better generation ability than the CNN, we can obtain better results by mixing the CNN features and the transformer features.

H. Computational Efficiency

As we use a PS strategy to improve the efficiency of the DRT algorithm, we evaluate the computational efficiency of the proposed network and analyze the impact of this strategy. In addition, we also compare the efficiency of DRT with other state-of-the-art algorithms. All methods are trained and executed on a server with an Nvidia TITAN RTX GPU and an Intel Core i7-7700 CPU, and the average running time is reported in Table VI, where DRT^* refers to the DRT framework without applying the PS strategy. In comparison, the computational time of the DRT^* method is dramatically reduced by using the PS strategy, so that DRT can run at a speed close to those of other representative nonrigid registration algorithms. We can conclude that the PS strategy is vital for our algorithm.

I. Scopes for Future Study

While our DRT method overcomes the shortcomings of equal-sized or rectangular patch embedding in current transformer methods and improves accuracy of nonrigid medical image registration, there still exists scopes for applying it to various situations. Although the PS strategy can improve the efficiency of DRT, the computational burden of the algorithm will still increase as the number of structures in images increases. For pathological images and other images with many complex structures, the searching time of the DRSRE module may be too long to be practical. It will also be difficult to embed the method in micro-controlled systems which have limited computing resources. In our future study, we will focus on improving efficiency of the searching process in the DRT algorithm.

V. Conclusion

This article proposes a DRT framework for nonrigid medical image registration with orientation constraints. The core of the proposed algorithm is the deformable region-based embedding module, which makes up for the shortcomings of equal-sized or rectangular patch embedding in current methods by adaptively learning to split the input images into deformable structural regions with orientation constraints. By applying the DRT, the DRT-based registration method preserves the semantic consistency across the regions to calculate accurate transformation relationships and thus performs better than other state-of-the-art CNN-based and transformer-based nonrigid registration methods. We also explore a PS strategy that considerably improves the efficiency of the DRT algorithm and makes it practical in run time. With deformable region-based embedding, DRT generates complete, robust, and discriminative features for each structural region in an image. So theoretically, it is not only applicable to medical images but also to other images which have multiple structural regions (i.e., facial images and remote sensing topographic maps). In addition, as the DRSRE module can work as a plug-and-play module and can easily be incorporated into various ViTs, it can be utilized in other ViTs as well as other region-based downstream tasks to improve their performance, e.g., image segmentation and object detection. Further verification experiments for applying DRT in these fields will be presented in our future work.

References

[1] Y. Yan, D. Wang, G. Zhou, and Q. Chen, "Unsupervised anomaly segmentation via multilevel image reconstruction and adaptive attention-level transition," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 5015712-1–5015712-12, 2021.

[2] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin transformer U-Net for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 4005615-1–4005615-15, 2022.

[3] H. Wang, Y. Ji, K. Song, M. Sun, P. Lv, and T. Zhang, "ViT-P: Classification of genitourinary syndrome of menopause from OCT images based on vision transformer models," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 4008014-1–4008014-12, 2021.

[4] Z.-M. Chen, Q. Cui, B. Zhao, R. Song, X. Zhang, and O. Yoshie, "SST: Spatial and semantic transformers for multi-label image recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 2570–2583, 2022.

[5] D. Wang, H. Hu, and D. Chen, "Transformer with sparse self-attention mechanism for image captioning," *Electron. Lett.*, vol. 56, no. 15, pp. 764–766, Jul. 2020.

[6] T. Liu et al., "Spatial-aware texture transformer for high-fidelity garment transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 7499–7510, 2021.

[7] Y. Wang, H. Cheng, and M. Q. H. Meng, "Inertial odometry using hybrid neural network with temporal attention for pedestrian localization," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 7503610-1–7503610-10, 2022.

[8] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.

[9] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[10] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[12] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," 2020, *arXiv:2006.04139*.

[13] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[14] Z. Chen et al., "DPT: Deformable patch-based transformer for visual recognition," 2021, *arXiv:2107.14467*.

[15] J. Chen et al., "ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration," 2021, *arXiv:2104.06468*.

[16] D. Hu, "Fusing CNNs and transformers for deformable medical image registration," in *Proc. 2nd Int. Conf. Comput. Sci., Electron. Inf. Eng. Intell. Control Technol. (CEI)*, Sep. 2022, pp. 19–23.

[17] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 548–558.

[18] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[19] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Effciently bridging CNN and transformer for 3D medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Strasbourg, France, Sep. 2021, pp. 171–180.

[20] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnFormer: Interleaved transformer for volumetric segmentation," 2021, *arXiv:2109.03201*.

[21] C.-F. Chen, R. Panda, and Q. Fan, "RegionViT: Regional-to-local attention for vision transformers," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 33–52.

[22] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–2.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[24] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[26] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4438–4446.

[27] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.

[28] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[29] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," *Comput. Sci.*, vol. 1, pp. 1–14, Feb. 2014.

[30] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[31] F. Shao, Z. Liu, and J. An, "Feature matching based on minimum relative motion entropy for image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 5603712-1–5603712-12, 2022.

[32] G. Gene, V. Loan, and F. Charles, *Matrix Computations*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996, pp. 56–57.

[33] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.

[34] S. Lan, Z. Guo, and J. You, "Non-rigid medical image registration using image field in demons algorithm," *Pattern Recognit. Lett.*, vol. 125, pp. 98–104, Jul. 2019.

[35] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Non-parametric diffeomorphic image registration with the demons algorithm," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2007, pp. 319–326.

[36] D. W. Shattuck et al., "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, Feb. 2008.

[37] S. G. Mueller et al., "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI)," *Alzheimer's Dementia*, vol. 1, pp. 55–66, Jul. 2005.

[38] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Med. Image Anal.*, vol. 5, no. 2, pp. 143–156, Jun. 2001.

[39] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters, "3D statistical neuroanatomical models from 305 MRI volumes," in *Proc. IEEE Conf. Rec. Nucl. Sci. Symp. Med. Imag. Conf.*, Oct./Nov. 1993, pp. 1813–1817.

[40] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 143–150, Feb. 2000.

[41] B. Fischl, "Freesurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[42] B. Avants, C. Epstein, M. Grossman, and J. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, Feb. 2008.

[43] M. Modat et al., "Fast free-form deformation using graphics processing units," *Comput. Methods Programs Biomed.*, vol. 98, no. 3, pp. 278–284, Jun. 2010.

[44] M. Heinrich, O. Maier, and H. Handels, "Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities," in *Proc. Int. Symp. Biomed. Imag.*, 2015, p. 27.

[45] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, Feb. 2005.

[46] G. Balakrishnan, A. Zhao, M. R. Sabuncu, A. V. Dalca, and J. Guttag, "An unsupervised learning model for deformable medical image registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9252–9260.

[47] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, Sep. 1997.

[48] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Med. Image Anal.*, vol. 57, pp. 226–236, Oct. 2019.

[49] B. Kim et al., "CycleMorph: Cycle consistent unsupervised deformable image registration," *Med. Image Anal.*, vol. 71, pp. 102–136, Jul. 2021.

[50] H. Qiu, C. Qin, A. Schuh, K. Hammernik, and D. Rueckert, "Learning diffeomorphic and modality-invariant registration using B-splines," *Med. Imag. With Deep Learn.*, vol. 143, pp. 645–664, 2021.

[51] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, no. 5, pp. 8024–8035.

[52] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, vol. 1, 2015, pp. 1–15.

[53] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.

[54] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image Anal.*, vol. 52, pp. 128–143, Feb. 2019.

[55] S. G. Madhuri and M. P. I. Gandhi, "Image registration quality assessment with similarity measures—A research study," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Apr. 2015, pp. 84–88.

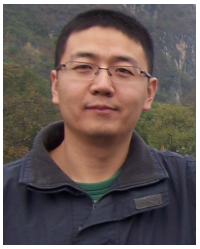
[56] H. Wang, X. Zhang, Y. Dou, and Y. Wang, "Edge connectivity evaluation algorithm for performance of edge detectors," *Comput. Eng. Appl.*, vol. 54, no. 16, pp. 192–196, 2018.

Sheng Lan received the M.S. degree in control engineering from Tsinghua University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree.

His main research interests are image processing, computer vision, biometrics, and medical image registration.

Xiu Li (Member, IEEE) received the Ph.D. degree in computer-integrated manufacturing from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2000.

From 2000 to 2002, she has served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China, where she has served as an Associate Professor from 2003 to 2010. Since 2016, she has been a Full Professor with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. Her research interests are in the areas of data mining, deep learning, computer vision, and image processing.



Zhenhua Guo received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, and the Ph.D. degree in computer science from The Hong Kong Polytechnic University, Hong Kong, in 2010.

Since April 2010, he has been working with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. He was a Visiting Scholar of electrical and computer engineering (ECE) with Carnegie Mellon University, Pittsburgh, PA, USA, from 2018 to 2019. His research interests include pattern recognition, computer vision, biometrics, and object detection.