MD-SGT: Multi-dilation spherical graph transformer for unsupervised medical image registration

Kun Tang, Lihui Wang, Xingyu Huang, Xinyu Cheng, Yue-Min Zhu

Highlights

MD-SGT: Multi-Dilation Spherical Graph Transformer for Unsupervised Medical Image Registration

Kun TANG,Lihui WANG,Xingyu HUANG,Xinyu CHENG,Yue-Min ZHU

- A multi-dilation graph transformer (MD-SGT) was proposed for unsupervised medical image registration to deal with the issues of the limited long-range spatial dependence and non-uniform attention spans in the existing methods.
- The features of each node of graph was updated by aggregating the information of its neighbors sampled from different spherical regions with different dilation rates.
- A group-wise convolutional layer instead of patch merging was used to downsample the graph for introducing the inductive bias of locality and transformation- equivariance into the graph transformer.
- The proposed MD-SGT outperforms the state-of-the-art registration methods, demonstrating that combining long-range uniform attention span and inductive bias are beneficial for promoting the image registration performance.

MD-SGT: Multi-Dilation Spherical Graph Transformer for Unsupervised Medical Image Registration

Kun TANG[a], Lihui WANG[a,*], Xingyu HUANG[a], Xinyu CHENG[a] and Yue-Min ZHU[b]

[a]*Engineering Research Center of Text Computing & Cognitive Intelligence, Ministry of Education, Key Laboratory of Intelligent Medical Image Analysis and Precise Diagnosis of Guizhou Province, State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, China*

[b]*University Lyon, INSA Lyon, CNRS, Inserm, IRP Metislab CREATIS UMR5220, U1206, Lyon 69621, France*

ARTICLE INFO

Keywords:
Deformable image registration
Convolutional neural network
Transformer
MD-SGT

ABSTRACT

Deformable medical image registration is an essential preprocess step for several clinical applications. Even though the existing convolutional neural network and transformer based methods achieved the promising results, the limited long-range spatial dependence and non-uniform attention span of these models prohibit further improving the registration performance. To deal with this issue, we proposed a multi-dilation spherical graph transformer (MD-SGT), in which the encoder combined the advantages of convolutional and graph transformer blocks to distinguish effectively the differences between the reference and the template images at various scales. Specifically, the features of each voxel were obtained by aggregating the information from its neighbors sampled from different spherical regions with different dilation rates. The implicit convolution inductive bias and long-range uniform attention span induced by such information aggregation manner made the features more representative for registration. Through the qualitative and quantitative comparisons with state-of-the-art methods on two datasets, we demonstrated that combining long-range uniform attention span and inductive bias are beneficial for promoting the image registration performance, with the Dice score, ASD and HD95 being improved at least by 0.5%, 2.2% and 1.1%, respectively.

1. Inroduction

Deformable image registration (DIR) intends to align the anatomical structures of two or more images by estimating the optimal transformations between them. DIR is a fundamental processing step for various clinical applications, such as image-guided treatment plan, prognosis evaluations, and disease monitoring etc. Although traditional registration methods (Avants et al. (2008); Modat et al. (2010); Heinrich et al. (2013); Vercauteren et al. (2009); Beg et al. (2005); Klein et al. (2009)) have achieved satisfactory performance, its time-consuming iterative optimization process for each individual pair data hindered its real-time applications. Recent advances in deep learning allow inferring the spatial transformations between any image pairs with a well trained network, such simple forward process makes it run much faster than conventional registration methods.

Currently, the deep learning based registration methods can be divided into the supervised and unsupervised ones. In the supervised-learning based registration methods, the ground-truth transformations between the reference and template images are usually obtained with conventional registration algorithms or spatial augmentation. With such ground-truth transformations as the objective, Yang et al. (2016) modeled a fully convolutional network (FCN) with the U-Net like architecture to predict the deformation field between different brain magnetic resonance (MR) volumes and achieved promising performance. Cao et al. (2017) proposed a similarity-steered convolutional neural network (CNN) to predict the deformation fields between the paired patches, with the patch similarity as the auxiliary contentual cue to guide the learning process, they obtained better registration results on several brain datasets. Sokooti et al. (2017) presented a RegNet which took a wide variety of artificially simulated displacement vecotr fields (DVFs) as the learning target and designed a patch-based multi-scale CNN to infer the DVFs. Although RegNet can obtain a satisfactory performance on CT images, due to the limits in diversity of the synthesized DVFs, it is difficult to transfer to the other dataset. To address this issue, Uzunova et al. (2017) proposed a locality-based shape and appearance model to generate huge amounts of image pairs and corresponding realistic ground truth DVFs, and then used a FlowNet to predict DVFs between 2D brain or cardiac MR images. They demonstrated that their method outperformed CNNs trained using either ground-truth DVF generated by conventional registration methods or that obtained by randomly augmenting the dataset. Even though the supervised-learning based registration methods have achieved much better registration results than conventional methods along with shorter computation time, their registration performance is highly dependent on the reliable ground-truth deformation fields which are not trivial to obtain in practice.

The emerge of spatial transformer network (STN) (Jaderberg et al. (2015)) makes it possible to implement image registration with unsupervised learning models. Vos et al. (2017) proposed a DIRNet which incorporated STN into the CNN architecture for the first time to predict the control points of B-spline transformation in an unsupervised manner. Subsequently, numerous STN-based unsupervised registration models have been presented. For instance, the group of Balakrishna et al. proposed a VoxelMorph model and the corresponding variants (Balakrishnan et al. (2018);

*Corresponding author: e-mail: lhwang2@gzu.edu.cn;

Dalca et al. (2018); Balakrishnan et al. (2019)), which used a U-Net-like architecture to predict the deformation field, and used a STN module to deform the template images, in addition, to guarantee the smoothness of DVFs and topology-preserving property, a squaring and scaling layer (Arsigny et al. (2006)) was designed before the STN module. Inspired by VoxelMorph, various modified models have emerged. Zhao et al. (2019a) proposed the recursive cascaded networks (RCN) for unsupervised medical image registration. Using VTN (Zhao et al. (2019b)) or VoxelMorph as the base network, the template image is warped successively by cascading the base network several times. They demonstrated that recursive cascaded networks can improve significantly the registration performance. Kuang and Schmah (2019) proposed a new registration algorithm FAIM, which introduced a new regularization term into the VoxelMorph to reduce the foldings in the DVFs. Subsequently, Mok and Chung (2020) proposed a novel Laplacian pyramid network (LapIRN) for deformable image registration, which mimics the traditional multi-resolution strategy of aligning image pairs from coarse to fine scale, and achieved promising registration results. Kim et al. (2021) proposed a CycleMorph network, which used cycle consistency to force the deformed image to return to the original image, such cycle consistency can enhance the image registration performance by preserving the original topology. Besides, Luo et al. (2021) proposed an adversarial registration network that contains two parts: one is the registrator for predicting deformation field, and another is the discriminator for determining whether two images' anatomical segmentation is well aligned, and their registration framework was proved to improve registration performance and training stability. Recently, Zhang et al. (2023) proposed a symmetric pyramid network for inverse consistent registration, which progressively conducts the feature-level diffeomorphic registration, and gained promising results against other registration methods. These unsupervised learning-based models have provided a promising mean to perform an end-to-end image registration quickly, however, most of the CNN-based models are not able to model the long-range dependency between the distant voxels. To deal with this problem, using atrous convolution instead of conventional convolution to enlarge the receptive field becomes an alternative (Devalla et al. (2018)). Nevertheless, the receptive fields of the low-level convolutional layers are still small, which are restricted by the kernel size. Moreover, Li et al. (2021) argued that the influence of distant pixels decays rapidly when the CNN becomes deeper. Therefore, the actual receptive fields of CNN are much smaller than the theoretical ones, which will hinder the CNNs from capturing differences between distant anatomies.

Considering the superiority of transformer in capturing the long-range spatial relationships, several DIR frameworks combining the transformer and CNN were proposed. For example, Chen et al. (2021) proposed a ViTVNet to perform 3D brain MR image registration, which incorporated ViT (Dosovitskiy et al. (2020)) into the VNet (Milletari et al. (2016)) model, enabling the model not only to capture the long-range spatial information but also to extract multi-scale representations. However, due to the high computational complexity and huge amounts of parameters, the transformer block can only be applied on the feature maps with small size. To solve this problem, Ma et al. (2022) proposed a symmetric transformer-based model (SymTrans), in which the convolution-based multi-head self-attention was proposed to reduce the parameters of the vanilla transformer, it achieved promising registration results with high computational efficiency. In the same year, the research group of the ViTVNet presented a hybrid transformer-CNN framework, TransMorph (Chen et al. (2022)), in which, the shifted window attention (Swin-transformer (Liu et al. (2021))) and patch merging were used in the encoder to capture the spatial correspondence between the reference and template images, and CNN served as a decoder to infer the DVFs from the output features of the encoder. To maintain the localization information, the long skip connections between encoder and decoder were deployed. They demonstrated that such hybrid framework can further improve the registration performance. Currently, shifted window self-attention is the most popular way to reduce the computational complexity of the vanilla transformer. However, as indicated by the work of neighborhood attention transformer (Hassani et al. (2022)), the attention span of each pixel (token or node) in Swin-transformer is not uniform, especially for the corner pixels in each window, which will influence the registration performance.

The non-uniform attention span mentioned above is mainly caused by the fixed window partition method, that means once the window was partitioned, for any pixel in this window, no matter where it locates, its features can be only updated by aggregating the information of other pixels limited in this window. To deal with this issue, the simplest way is to update the features of a given pixel using the information of pixels sampled from its self-centered regions. In the image space, window-variant self-attention is not easy to implement. Considering the flexibility of graph neural network in aggregating the neighboring information, such as in graph convolutional neural network (GCN) Kipf and Welling (2016), it updates the features of a given node by aggregating the transformed information of other graph nodes through the adjacent matrix; in graph attention network (GAT) Veličković et al. (2017), it updates the features of a target node by aggregating the attention-weighted information of its one-order neighbors. Once the neighborhood in the graph space is defined appropriately and uniformly, the problem of non-uniform attention span can be easily solved. Inspired by this idea, we proposed a multi-dilation spherical graph Transformer (MD-SGT) for unsupervised DIR. Specifically, for each pixel, we first sampled its neighbors with different dilation rates from multiple spherical regions defined by different radii, and then updated the features of each pixel by aggregating the information

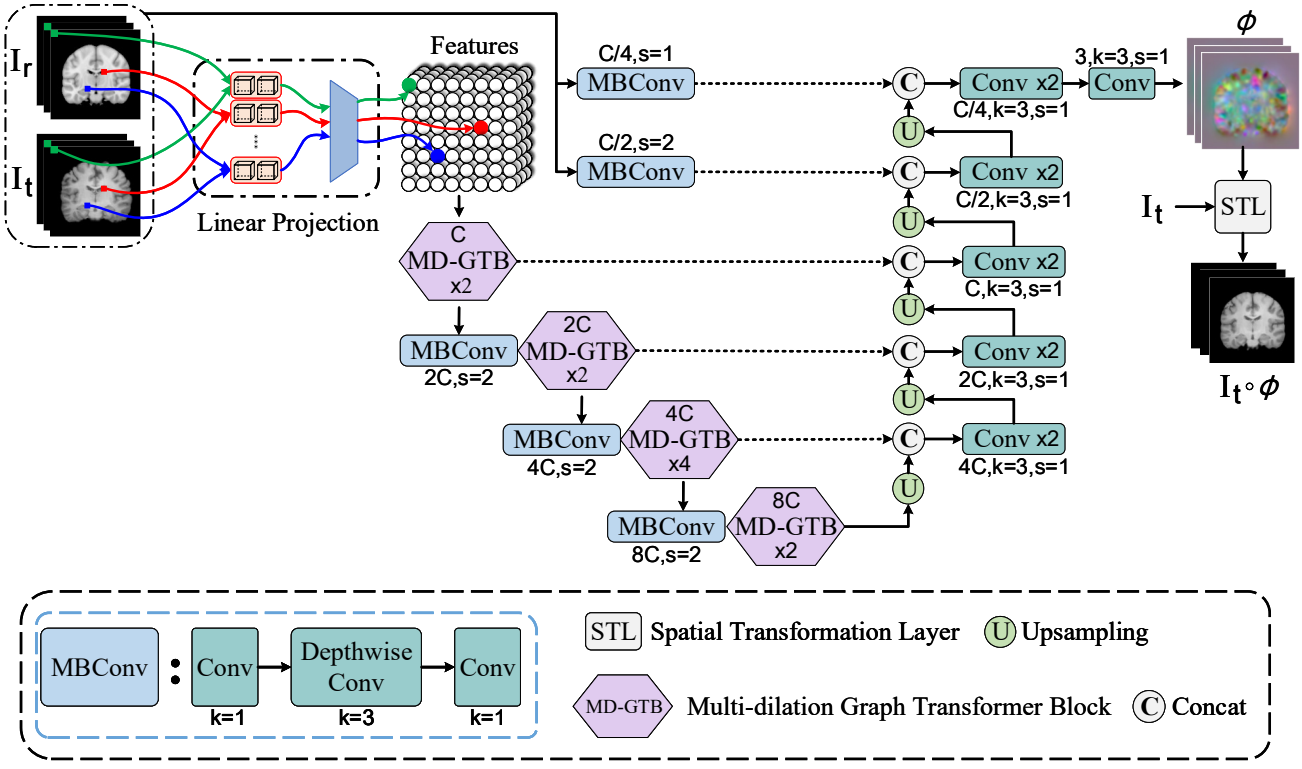


Figure 1: The architecture of MD-SGT network for image registration. MD-SGT has an encoder-decoder architecture, its encoder is composed by a series of MD-GTB and MBConv modules, where MD-GTB is responsible to extract the useful features using graph self-attention operations, and MBConv is used for downsampling the feature maps; its decoder is simply composed by several successive upsampling and convolutional layers.

from its neighbors with graph self-attention. Such multi-scale sampling and information aggregation mechanism allows each pixel having a uniform and long range attention span. At the same time, instead of the patch merging, we used a group-wise convolutional layer to downsample the nodes so that the inductive biases of the convolution, including the locality and translation equivariance, can be introduced into the graph transformer to further improve the registration performance.

In the rest of this paper, the details of the proposed MD-SGT are presented in section 2. The dataset along with the preprocessing, and the experimental implementation details are described in Section 3. Both qualitative and quantitative comparisons as well as the ablation results are demonstrated in Section 4. Finally, a thorough analysis about the results and the limitations of this work are discussed in Section 5, followed by a conclusion in Section 6.

2. Method

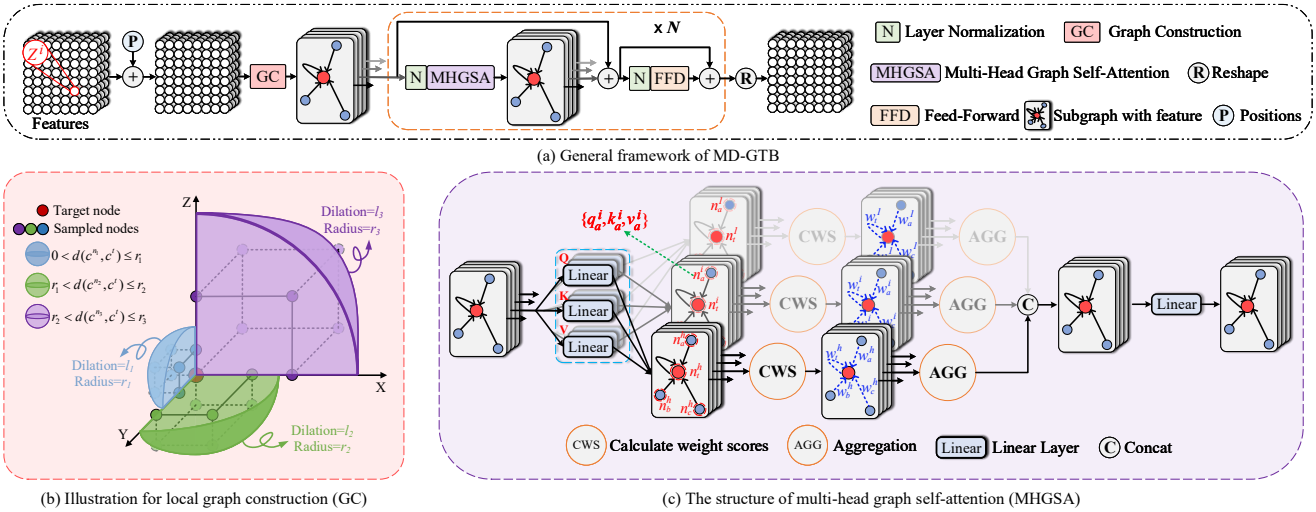
Given the reference image I_r, and the template I_t which needs to be deformed, the task of image registration is to establish the anatomical correspondences between I_r and I_t. To achieve this, we proposed a multi-dilation spherical graph Transformer (MD-SGT) to learn the deformation field, as illustrated in Figure 1(a). MD-SGT uses an encoder composed with several convolution blocks and multi-dilation graph transformer blocks (MD-GTB) to extract the semantic features of both I_r and I_t, as well as a decoder composed with several upsampling layers and convolution layers to infer the deformation field between I_r and I_t. Finally, the spatial transformation (ST) layer is applied for warping the template image I_t with deformation field. In the following subsections, the structure of MD-GTB, the principle of ST layer and the loss functions will be elaborated in detail.

2.1. Structure of MD-SGT

As illustrated in Figure 1, the reference image I_r and template image I_t are firstly split into the paired patches with a specific size of $p \times p \times p$. Subsequently, a linear embedding layer is used to project each paired patch into a d-dimensional feature vector. Denoting the paired image patches as $p^i \in \mathbb{R}^{(2 \cdot p \cdot p \cdot p) \times 1}$, with $i \in 1, ..., N$ and N being the number of the patch pairs, which is equal to $(H \times W \times D)/p^3$ if the image size is $H \times W \times D$, the resulting feature vector $z \in \mathbb{R}^{N \times d}$ for each pair of patches can be written as

$$z^i = M \otimes p^i. \tag{1}$$

where $M \in \mathbb{R}^{d \times (2 \cdot p \cdot p \cdot p)}$ represents the learnable parameter matrix of a linear projection layer and $\otimes$ indicates the matrix multiplication. Each paired patches expressed by a d-dimensional feature vector can be taken as a node distributed on a regular grid graph. To capture effectively



(a) General framework of MD-GTB

(b) Illustration for local graph construction (GC)

(c) The structure of multi-head graph self-attention (MHGSA)

Figure 2: Detailed structure of MD-GTB module. The general framework of MD-GBT is given in (a). It mainly consists of the graph construction (GC) module (b) and multi-head graph self-attention (MHGSA) module (b). GC shows how to construct the local graph form the input images and MHGSA illustrates the process of graph self-attention calculation.

and efficiently the correlations between different nodes, the node features are input into the MD-GTB modules (Figure 2(a)), in which they pass through firstly a positional encoding layer to generate a specific position vector pos^i for each node i. The dimension of pos^i is the same as that of node feature vector z^i. After adding the positional vector, the feature of each node becomes

$$z^i = z^i + pos^i. \tag{2}$$

Subsequently, for each node, a local graph is constructed by sampling its spherical neighboring nodes with different dilation rates and radii. Specifically, let Ω_R be a set of coordinates in a region of a radius of R around a given target node t. Here, R is much larger than the sampling radius r. Sampling the nodes with a dilation rate of l from Ω_R, with the coordinates denoted as $c^n = (c^t + k \times l) \in \Omega_R$ (k is an integer larger than 0), if the distance between the target node and the sampling node is less than the sampling radius r, such node is considered as a neighbor:

$$n \in N\{t\} \quad \text{if} \quad d(c^n, c^t) \le r \tag{3}$$

where $d(c^n, c^t)$ represents the Euclidean distance between the target node t and sampling node n, $N\{t\}$ represents the set of neighboring nodes of t. In this work, we used three different dilation rates to sample the possible neighboring nodes within three radii, to avoid repeated sampling, the neighbors of target node t are selected with:

$$\begin{aligned} n_1 &\in N\{t\} \quad \text{if} \quad d(c^{n_1}, c^t) \le r_1 \\ n_2 &\in N\{t\} \quad \text{if} \quad r_1 < d(c^{n_2}, c^t) \le r_2 \\ n_3 &\in N\{t\} \quad \text{if} \quad r_2 < d(c^{n_3}, c^t) \le r_3 \end{aligned} \tag{4}$$

where r_j and l_j are the j^{th} $(j = 1, 2, 3)$ sampling radius and dilation rate, respectively. The coordinates of possible sampling nodes $c^{n_j} = (c^t + k \times l_j) \in \Omega_R$. The idea of local graph construction can be seen in Figure 2(b).

Using such local graph, the feature of each target node can be updated through multi-head graph self-attention (MHGSA) module in Figure 2(c). Specifically, for each target node t, its query q^t, key k^t, and value v^t vectors can be obtained through fully connection layers,

$$\begin{aligned} q^t &= Norm(z^t) \cdot w_q \\ k^t &= Norm(z^t) \cdot w_k, \\ v^t &= Norm(z^t) \cdot w_v \end{aligned} \tag{5}$$

where w_q, w_k, w_v are three linear transformation matrices with size of $d \times d$, $Norm(\cdot)$ indicates the layer normalization (Ba et al. (2016)), and $\cdot$ means dot product. Accordingly, the weight score of the neighboring node i for the target node t is calculated by:

$$w^i = \exp\left(\frac{q^t \cdot k^i}{\sqrt{d}}\right) \tag{6}$$

where k^i is the key vector of node i, also calculated with (5). According to the self-attention mechanism, the information of all the neighboring nodes are aggregated on the target node and then pass through a linear transformation to derive the output of MHGSA,

$$o^t = \frac{\sum_{i \in N(t)} w^i \cdot v^i}{\sum_{i \in N(t)} w^i} \cdot w_o, \tag{7}$$

where $N(t)$ is the neighboring nodes of target node t, v^i is the value vector of the neighboring node i, and the w_o is the parameter matrix of the linear transformation layer with dimension of $d \times d$. As illustrated in Figure 2(c), after passing through two residual operations, layer normalization and feed forward (FFD) layer, the final feature vector o^{t_f} of a target node t is formulated as:

Algorithm 1: Performing self-attention on graph

Data: Input feature $z^t \in \mathbb{R}^d$ for each node t in pre-defined graph

for *each node t in graph* **do**

Calculate Query q^t, Key k^t, Value v^t by $q^t = Norm(z^t) \cdot w_q, k^t = Norm(z^t) \cdot w_k, v^t = Norm(z^t) \cdot w_v$;

Calculate weight score w^i of each neighbor linked with t by $w^i = exp(\frac{q^t \cdot k^i}{\sqrt{d}})$;

Aggregate the information of neighbors on t by $o^t = \frac{\sum_{i \in N(t)} w^i \cdot v^i}{\sum_{i \in N(t)} w^i} \cdot w_o$;

Update node feature o^t with the following operations: $o^{tmp} = z^t + o^t$; $o^{t\_f} = o^{tmp} + FFD(Norm(o^{tmp}))$

end

$$\begin{aligned} o^{tmp} &= z^t + o^t \\ o^{t\_f} &= o^{tmp} + FFD(Norm(o^{tmp})). \end{aligned} \tag{8}$$

Updating the features of all the nodes will generate the output of MD-GTB, the detailed process is given in the Algorithm 1. As demonstrated in Figure 1, after each MD-GTB group, motivated by the MaxViT (Tu et al. (2022)), the grid nodes are down-sampled with a MBConv layer with stride of 2. In addition, the radii and dilation rates used in MD-GTB modules (from top to down) are set as $3^1 6^2 9^3$, $3^1 6^2 9^3$, $3^1 6^2 9^3$ and 8^1, respectively, in which $3^1 6^2 9^3$ indicates that the neighboring nodes are sampled from three spherical regions with radii of 3, 6 and 9 respectively, the corresponding dilation rates are set as 1, 2 and 3. For the last MD-GTB module, the field of view (FOV) of grid nodes is so small that all the nodes can be taken as the neighbors, accordingly, the sampling radius is 8 and the dilation rate is 1.

Since the MD-GTB modules used in the encoder can capture the long-range dependence between two distant voxels, and the first two MBConv layers in the encoder can extract the local information, transmitting both local and long-range information into the decoder is beneficial for deriving the deformation field more accurately.

2.2. Spatial transformation layer

The deformation field and the template image are input into the spatial transformation layer to perform coordinate transformation and interplolation. In other words, defining the template image intensity at location of p as $I_t(p)$, the output of the deformed image at the same location is defined by:

$$I_t \circ \phi(p) = I_t(p + \phi(p)) \tag{9}$$

The values of deformation field are continuous, the deformed location $p + \phi(p)$ could be not integer, the interpolation is therefore required to calculate the intensity $I_t(p + \phi(p))$, that means,

$$I_t(p+\phi(p)) = \sum_{q \in N(p+\phi(p))} I_t(q)(1 - d(p + \phi(p), q)) \tag{10}$$

where $N(p + \phi(p))$ includes eights neighboring voxels around $p+\phi(p)$, and $d(p+\phi(p), q))$ represents the Euclidean distance between the location $p + \phi(p)$ and its neighboring point q.

2.3. Loss function

The loss function used in this work is constituted of two parts, as demonstrated in (11), $L_{sim}(\cdot)$ is for measuring the dissimilarity degree between the reference image I_r and the warped image $I_t \circ \phi$, and $L_{reg}(\cdot)$ is for guaranteeing the smoothness of the deformation field ϕ, formulated as:

$$L(I_r, I_t, \phi) = L_{sim}(I_r, I_t \circ \phi) + \alpha * L_{reg}(\phi) \tag{11}$$

where α is the trade-off parameter used to balance the $L_{sim}(\cdot)$ and $L_{reg}(\cdot)$. In this work, the negative normalized cross-correlation (NCC) is used as $L_{sim}(\cdot)$ loss, written as:

$$\begin{aligned} &L_{sim}(I_r, I_t \circ \phi) = -NCC(I_r, I_t \circ \phi) = \\ &-\sum_{p \in \Omega} \frac{(\sum_{p_i} (I_r(p_i) - \bar{I}_r(p))(I_t(p_i + \phi(p_i)) - \bar{I}_t(p + \phi(p))))^2}{\sqrt{\sum_{p_i} (I_r(p_i) - \bar{I}_r(p))^2 \sum_{p_i} (I_t(p_i + \phi(p_i)) - \bar{I}_t(p + \phi(p)))^2}} \end{aligned} \tag{12}$$

where Ω is the set of image voxels, p_i represents any voxels inside the neighborhood $N(p)$ of a given voxel p, $\bar{I}_r(p)$ and $\bar{I}_t(p + \phi(p))$ indicate the mean intensity value of the local region around p in the reference and warped image, respectively. As to the regularization loss $L_{reg}(\cdot)$, we use the diffusion regularizer to encourage the deformation field to be smooth,

$$L_{reg}(\phi) = \sum_{p \in \Omega} \left\| \partial \phi_x(p) \right\|^2 + \left\| \partial \phi_y(p) \right\|^2 + \left\| \partial \phi_z(p) \right\|^2 \tag{13}$$

where $\partial \phi_x(p)$, $\partial \phi_y(p)$, and $\partial \phi_z(p)$ are the spatial gradients of ϕ along x, y, z axis respectively.

3. Experiments

Several comparison and ablation experiments are implemented to evaluate the performance of the proposed MD-SGT network. Particularly, we compare MD-SGT with state-of-the-art registration methods including Affine, NiftyReg (Modat et al. (2010)), deedsBCV (Heinrich et al. (2013)), MIDIR (Qiu et al. (2021)), Recursive Cascaded Networks (Zhao et al. (2019a)), CycleMorph (Kim et al. (2021)), VoxelMorph (Balakrishnan et al. (2018)), LapIRN (Mok and Chung (2020)), ViTVNet (Chen et al. (2021)), TransMorph (Chen et al. (2022)) and symTrans (Ma et al. (2022)) on two datasets. The description of the datasets, the implementation details of all the models, the ablation settings and the quantitative evaluation metrics will be elaborated in the following subsections.

3.1. Dataset and Preprocessing

In this work, two datasets corresponding to different registration tasks, Atlas-to-Patient and Patient-to-Atlas registrations, were used.

3.1.1. Atlas-to-Patient brain MRI registration dataset

In Atlas-to-Patient registration, a public available dataset was used, where 576 T1-weighted (T1w) MR brain images in IXI dataset[1] provided by Transmorph (Chen et al. (2022)) were used as references (Patient), and the atlas was provided by CycleMorph (Kim et al. (2021)). All scans were preprocessed with the FreeSurfer (Fischl (2012)) and cropped to size of $160 \times 192 \times 224$. Consistent with the Transmorph, 403, 58, and 115 scans were used as training, validation, and test sets, respectively.

3.1.2. Patient-to-Atlas brain MRI registration dataset

For Patient-to-Atlas registration, we used the in-house dataset, including 3D T1w MR images of 102 drug-addicts and 10 healthy controls, which were acquired from Guizhou Provincial People's Hospital using a 3.0T MRI scanner (GE 3.0T Discovery 750W) with a 32 channel head and neck coil. The imaging parameters are: repetition time (TR) = 8.464 ms, echo time (TE) = 3.248 ms, inversion time = 450 ms, flip angle = 15°, field of view (FOV) = 256×256 mm^2, matrix $=256\times256$, slice thickness = 1.0 mm, and slice gap = 0 mm. Additionally, the MR images were acquired in the sagittal plane, yielding 188 continuous slices, with a resolution of $1.0 \times 1.0 \times 1.0$. The skull was removed firstly with BET method (Smith (2002); Jenkinson et al. (2005)) embedded in FSL, and then the image intensity was normalized to the range [0, 255], finally, the normalized images were linearly aligned with the MNI152 template provided by the McConnell brain imaging centre using FSL FLIRT method (Jenkinson and Smith (2001); Jenkinson et al. (2002)), and then resampled to the size of $128 \times 128 \times 128$.

3.2. Implementation details

The proposed method was implemented with PyTorch (Paszke et al. (2019)) and DGL (Wang et al. (2019)). All the learning-based comparison methods were trained on a GPU of NVIDIA A100 (40GB) for 500 epoches using Adam optimizer (Kingma and Ba (2014)), with batch size of 1 and learning rage of 0.0001. The detailed settings for each method were given as follows:

1. NiftyReg (Modat et al. (2010)): To achieve its best registration performance, the locally normalized cross-correlation (LNCC) was replaced with the normalized mutual information (NMI) as the objective function, and bending energy was used as the regularization with a weighting coefficient of 0.001. The iteration number was 300.
2. deedsBCV (Heinrich et al. (2013)): The default settings of deedsBCV were used in this work, in which the objective function was set as the self-similarity context (SSC), the weight of smooth regularization was set to be 0.4, and 5 scale levels with grid spacing changing from 8 to 4 voxels for B-spline interpolations were used.
3. MIDIR (Qiu et al. (2021)): The objective function of MIDIR consisted of NCC and the L2 regularization item (weight was set as 1) . The spacing of control points used for B-Spline interpolation was 3.
4. Recursive Cascaded Networks (RCN) (Zhao et al. (2019a)): Default similarity measurement and regularizations were used. Since the reference and template images used in this work have already been aligned with affine registration, thus in our implementations, the affine registration module of RCN was removed, and voxelMorph was selected as the basic cascade network (cascade number was set as 5).
5. VoxelMorph (Balakrishnan et al. (2018)): VoxelMorph-1 with the default settings was used in this work.
6. CycleMorph (Kim et al. (2021)): The weighs of the registration loss, cycle loss, identity loss and regularization were set as 1, 0.1, 0.5, and 1, respectively, to get the best performance.
7. LapIRN (Mok and Chung (2020)): Default settings of the LapIRN were used in this work and the number of levels was set as 3.
8. ViTVNet (Chen et al. (2021)): The default settings of the ViTVNet architecture were used in this paper, in which the patch size was $8 \times 8 \times 8$, the number of heads and the number of transformer layers were 12. As to the loss function, NCC followed with L2 regularization (weight was 1) was used.
9. TransMorph (Chen et al. (2022)): The same loss function and weight as the VoxelMorph were used. The patch size was 4, the window size of the Swin-Transformer was 8, and numbers of heads for each level were 4, 4, 8, and 8, respectively.
10. SymTrans (Ma et al. (2022)): The default settings of SymTrans were used.
11. Proposed model: The patch size was 8 and the weight of regularization item α in Eq. (11) was 1.

3.3. Ablation studies

In the ablation studies, we investigated the effects of several factors on the registration performance, detailed as follows.

3.3.1. Effectiveness of MD-GTB and MBConv modules

To validate the effectiveness of the MBConv and the proposed MD-GTB modules, we performed several ablation studies. Specifically, we adopted the VoxelMorph-like architecture as the baseline model (indicated by Base in the Table 1), in which the MaxPooling along with one convolutional layer rather than MBConv or MD-GTB module was used for downsampling feature maps in the encoder. In addition, to make the fair comparison, the decoder was kept the same as ours, and the number of channels for all the layers was consistent with the proposed model. Based on

[1] https://brain-development.org/ixi-dataset/

Table 1
Effects of MBConv and MD-GTB modules.

Model	MBConv	MD-GTB
Base	✗	✗
Base+MBConv	✓	✗
Base+MD-GTB	✗	✓
Base+MBConv+MD-GTB	✓	✓

Table 2
Influence of the number of MD-GTB modules and the attention heads

Model	Embed. Dimension	MD-GTB numbers	Head numbers
MD-SGT	96	{2, 2, 4, 2}	{4, 8, 16, 32}
MD-SGT small	48	{2, 2, 4, 2}	{4, 8, 16, 32}
MD-SGT large	128	{2, 2, 12, 2}	{4, 8, 16, 32}
MD-SGT-NH1	96	{2, 2, 4, 2}	{1, 1, 1, 1}
MD-SGT-NH4	96	{2, 2, 4, 2}	{4, 4, 4, 4}
MD-SGT-NH8	96	{2, 2, 4, 2}	{8, 8, 8, 8}

such baseline model, we replaced all the MaxPooling layers in the encoder with the MBConv (Base+MBConv) and MD-GTB (Base+MD-GTB) respectively, we then compared these three models with our proposed method.

3.3.2. *Effects of the model hyperparameters*

The effects of model hyperparameters, including the number of MD-GTB modules in each level of the encoder, the number of attention heads in each MD-GTB module, as well as the embedding dimension, on registration performance were also investigated in this work. As demonstrated in the top panel of Table 2, we varied the number of MD-GTB modules in the fourth level of encoder from 4 to 12 and the embedding dimension from 96 to 128 to show the performance difference between the small (MD-SGT small), medium (MD-SGT) and large (MD-SGT large) models. In addition, we changed the number of attention heads in the MD-SGT model to show its influence, as listed in the bottom panel of Table 2.

3.3.3. *Influence of the Radius of Neighborhood*

As illustrated in Figure 2(b), during the graph construction, the selection of radii of multi-dilation neighborhoods will influence the number of neighbors of a given node and then further affect the registration performance. Accordingly, based on the MD-SGT model architecture, we have varied the radius of the sampling area for different dilation rates at three scales (except the last scale where the FOV is so small that all the nodes are considered as the neighbors.), as listed in Table 3, at each scale, corresponding to the dilation rates of (1, 2, 3), the sampling radius changed from (1, 7, 9) to (2, 5, 10) and (3, 6, 9). For the last scale, the radius and dilation rate were always 8 and 1, respectively.

3.3.4. *Influences of Neighbor Sampling Schemes*

The neighbor sampling scheme proposed in MD-GTB module can allow the graph transformer to have a uniform attention span. To verify whether the uniform attention span has the advantages in DIR, we first sampled the neighbors for any target nodes in a fixed sphere (non-uniform attention span) and also sampled the neighbors of a given target node from self-centered sphere (uniform attention span), as illustrated in Figure 3 (a). After that, to further validate the effectiveness of the multi-dilation neighboring sampling scheme in MD-GTB, we replaced the multi-dilations with a single dilation, denoted as SD-GTB block. It means that, during the graph construction, the neighboring nodes of one target are only sampled from one sphere with fixed radius, as illustrated in Figure 3 (b). In the comparison, all the MD-GTB blocks in the proposed model applied at four different scales are replaced with SD-GTB blocks, to avoid the influence of the number of neighboring nodes, we set the radii of spherical neighborhood of SD-GTB and MD-GTB as $\{4^1, 4^1, 4^1, 8^1\}$ and $\{3^1 6^2 9^3, 3^1 6^2 9^3, 3^1 6^2 9^3, 8^1\}$ respectively, making sure that the number of neighbors in MD-GTB and SD-GTB being almost the same. Finally, considering that our multi-dilation sampling strategy is similar to atrous convolution, we also compared the performance of MD-GTB module and atrous convolution by replacing the MD-GTB modules in the proposed network with atrous convolutions. For the fair comparison, we let the atrous convolution and MD-GTB module have the same receptive field size.

Table 3
Effects of neighborhood radius in MD-GTB.

Model	Radii & Dilations
MD-SGT-179	$\{1^1 7^2 9^3; 1^1 7^2 9^3; 1^1 7^2 9^3; 8^1\}$
MD-SGT-2510	$\{2^1 5^2 10^3; 2^1 5^2 10^3; 2^1 5^2 10^3; 8^1\}$
MD-SGT-369	$\{3^1 6^2 9^3; 3^1 6^2 9^3; 3^1 6^2 9^3; 8^1\}$

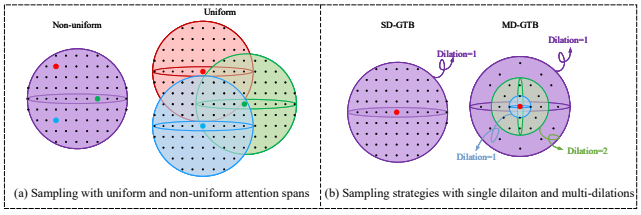


Figure 3: Illustration of different neighbor sampling schemes. Sampling the neighboring nodes with uniform and un-uniform attention spans (a), as well as sampling the same amount of neighbors with single and multiple dilation rates (b), respectively.

3.3.5. *Effects of Graph Attention Manners*

Given a target node, its neighboring nodes are sampled from different field of views (FOVs), accordingly, the attentions between the target node and neighboring nodes can be calculated in different ways. The first one is to calculate the attention maps between the target nodes and all the neighboring nodes in different FOVs (Figure 4(a)), and then the features of the target node are updated with attention-weighted neighboring information, we called this manner as

"integration". The second one is to calculate the attention maps at different FOVs respectively, it means that we update the features of target nodes with the different neighboring nodes for three times and then combine these three target features through a convolution layer (Figure 4(b)), such manner was named after "parallelly split". The last one is also to calculate the attention maps at different FOVs respectively, but the features of target nodes are updated serially rather than parallelly, which means the features of target node are updated firstly with the neighboring nodes in FOV3, such target features will be further updated with the neighboring nodes in FOV2, as illustrated in Figure 4(c). In this work, such manner was called as "serially split".

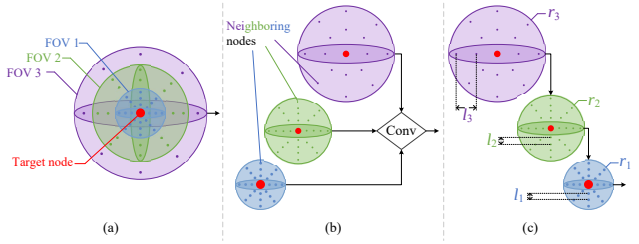


Figure 4: Illustration of different attention manners. (a) Integration manner, (b)Parallely split manner, (c) Serially split manner.

3.4. Evaluation metrics

To quantitatively evaluate the registration performance of each method, dice score (Dice), average surface distance (ASD), and the Hausdorff distance (95%, HD95) were calculated. Dice measures the overlap between two segmented anatomical regions A_{seg} and B_{seg}:

$$Dice = 2 \times \frac{\left|A_{seg} \cap B_{seg}\right|}{\left|A_{seg}\right| + \left|B_{seg}\right|} \tag{14}$$

The maximum value of Dice is 1, and larger Dice indicates the higher overlap between the two regions.

ASD and HD are boundary-based measurements which evaluate the closeness in boundaries between two regions, defined as:

$$ASD = \frac{\sum_{x \in \partial B_{seg}} d(x, \partial A_{seg}) + \sum_{y \in \partial A_{seg}} d(y, \partial B_{seg})}{\left|\partial A_{seg}\right| + \left|\partial B_{seg}\right|} \tag{15}$$

$$HD = \max\{\max_{x \in \partial B_{seg}} d(x, \partial A_{seg}), \max_{y \in \partial A_{seg}} d(y, \partial B_{seg})\} \tag{16}$$

where ∂A_{seg} and ∂B_{seg} represent the gradient map of A_{seg} and B_{seg}, respectively, x and y indicate an arbitrary voxel in the gradient maps ∂A_{seg} and ∂B_{seg}, $d(x, \partial A_{seg})$ and $d(y, \partial B_{seg})$ are the corresponding minimum Euclidean distance between x/y and set of surface voxels of $\partial A_{seg}/\partial B_{seg}$. Lower ASD or HD95 indicates better registration performance.

In addition, we also reported the percentages of non-positive values in the determinant of the Jacobian matrix on the deformation fields to quantify the regularity of the deformation fields.

4. Results

4.1. Atlas-to-Patient registration results with IXI dataset

The qualitative results of atlas-to-patient registration on IXI dataset obtained with different methods were given in Figure 5, in the first column of which contains the reference image (Ref.) and the linearly aligned template image (Aff.), and in the rest columns, the warped images derived from different methods, along with corresponding deformation field covered by deformed grid were illustrated. As shown in the zoomed-in areas, the NiftyReg and the deedsBCV produce erroneous shapes and over-smoothed deformation fields, indicating that they cannot infer accurately the deformations around this region. In other words, the two methods can only roughly align two images but not the small structures. In contrast, transformer-based methods can achieve better alignment results, especially for the proposed MD-SGT method which generates a warped image most similar to the Ref. The performance of the CNN-based methods (except CycleMoprh) is not better than deedsBCV on IXI dataset, especially for the method of LapIRN, which achieves the lowest Dice score. Moreover, as we can observe in the deformed grid, the phenomenon of folding in deformation field (cyan arrows) occurs in all the learning-based methods, but the proposed method has the fewest folding, indicating the superiority of proposed MD-SGT on aligning two images along with more plausible deformations.

The quantitative evaluation metrics for atlas-to-patient registration are listed in Table 4. In each row shows the mean and standard deviation of the Dice, ASD, and HD95 of each method in 29 anatomical structures for all the images in the test set. We observed that the proposed method MD-SGT achieves the highest mean Dice score and the second lowest ASD and HD95 distances on IXI dataset. Comparing with the state-of-the-art (SOTA) methods, TransMorph and SymTrans, the Dice score, ASD and HD95 distance are improved by 0.5% / 0.8%, 2.2% / 3.1%, and 1.1% / 3.6% respectively. Even though the learning-based methods (except VoxelMorph) have better performance than traditional methods (NiftyReg and deedsBCV), the folding effects in deformation field are more serious (i.e., higher values for % of $|J_\phi|$). As shown in Table 4, the mean determinant of the Jacobian matrix on the deformation fields ($|J_\phi|$) of our method is the smallest one among the learning-based methods, improved at least 4% comparing with the SOTA methods. Such superiority can also be found in Figure 6, where the curves of Dice score, ASD and HD95 for each sample in test set are drawn. We notice that, among all the learning-based methods, for almost all the samples, our method (purple curves) produced the highest Dice scores and the second lowest ASD and HD95 distances.

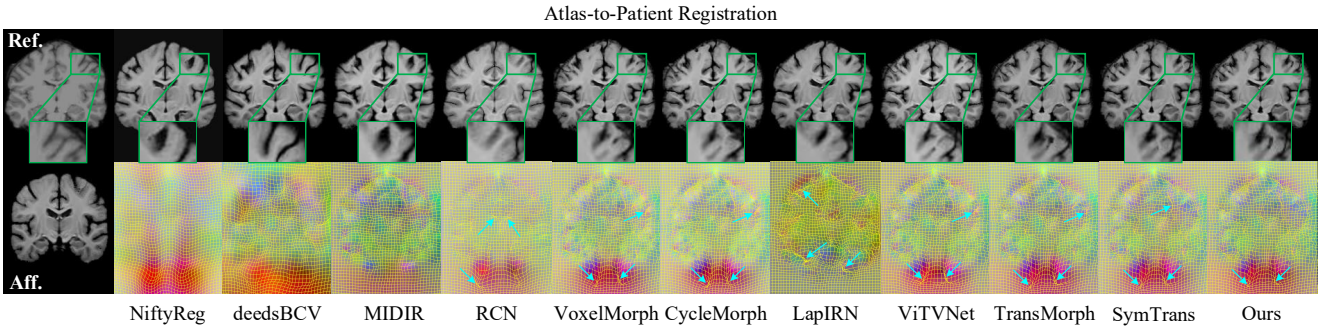


Figure 5: Qualitative comparison between the reference and the warped images obtained with different methods on the IXI dataset. The first column contains the reference image (Ref.) and the linearly aligned template image (Aff.), and each of the rest columns contains the warped image along with the corresponding deformation field covered by deformed grid. The cyan arrows indicate the folding in the deformation field.

Table 4
Average DICE, ASD and HD95 scores for different methods performing atlas-to-patient registration.

	Atlas-to-Patient			
Method	Dice↑	ASD↓	HD95↓	% of $\lvert J_\phi \rvert \downarrow$
Affine	0.386±0.195	2.800±0.865	6.771±1.460	-
NiftyReg (Modat et al. (2010))	0.634±0.169	1.834±1.014	5.683±3.513	0.006±0.016
deedsBCV (Heinrich et al. (2013))	0.733±0.126	1.163±0.666	3.751±2.559	0.052±0.066
MIDIR (Qiu et al. (2021))	0.742±0.128	**1.050±0.465**	**3.352±1.936**	< 0.0001
RCN (Zhao et al. (2019a))	0.724±0.120	1.324±0.767	4.234±2.824	0.369±0.190
VoxelMorph (Balakrishnan et al. (2018))	0.729±0.129	1.179±0.626	3.967±2.468	1.573±0.336
CycleMorph (Kim et al. (2021))	0.737±0.123	1.165±0.654	3.918±2.595	1.701±0.378
LapIRN (Mok and Chung (2020))	0.625±0.160	1.621±0.580	4.662±1.611	2.061±0.672
ViTVNet (Chen et al. (2021))	0.734±0.124	1.162±0.603	3.939±2.453	1.590±0.316
TransMorph (Chen et al. (2022))	0.753±0.126	1.091±0.640	3.756±2.536	1.497±0.341
SymTrans (Ma et al. (2022))	0.751±0.127	1.102±0.649	3.854±2.599	1.673±0.350
MD-SGT (Ours)	**0.757±0.125**	1.068±0.635	3.714±2.574	1.392±0.342

To further evaluate the registration performance in different brain regions, Figure 7 shows the boxplots of Dice, ASD and HD95 obtained with different methods in 17 brain ROIs, including brain stem, thalamus, cerebellum cortex (CC), cerebral white matter (CWM), cerebellum white matter (CeWM), putamen, ventralDC, pallidum, caudate, lateral ventricle (LV), hippocampus, 3rd-ventricle, 4th-ventricle, amygdala, cerebral cortex (CeCo), CSF and choroid plexus (CP). We noticed that, the proposed method achieves the highest median Dice scores and the narrowest interquartile range in the most brain regions, indicating that the brain regions warped with our method can align well with the corresponding reference regions.

4.2. Patient-to-Atlas registration results with in-house dataset

To compare visually the performance of different methods on patient-to-atlas registration task, the registration results of one randomly selected slice in the test set of in-house data were demonstrated in the top of Figure 8, and the corresponding deformation grids as well as the displacement vector fields were shown in the bottom, where the red arrows indicated the irregular or folded deformation field. We observed that NiftyReg, deedsBCV, and MIDIR cannot deal with the shape change of the lateral ventricle well. Moreover, the conventional algorithms and most of the learning-based methods cannot account for the deformation in the area zoomed-in on the middle bottom, but generally, the learning-based methods generated better warped images than conventional methods, especially for our method and LapIRN. As highlighted in the green rectangles, the LapIRN can retain well the structure details in the warped image and without folding effect. Even though there are still some irregular deformations in the displacement vector filed of the proposed method (indicated by the red arrows), in contrast to the rest learning-based methods, the deformation field obtained by our method is much smoother.

The quantitative evaluation metrics for patient-to-atlas registration are given in Table 5, where the mean and standard deviation of the Dice, ASD and HD95 of each method in 58 anatomical regions for all individuals in the test set are demonstrated. We observed that, in the conventional methods, deedsBCV obtained the best registration results in terms of three evaluation metrics, which were even better than the learning-based method (MIDIR). In addition, the standard deviations of all the evaluation metrics for the conventional methods were much lower than the learning-based methods, indicating that the non-deep-learning-based

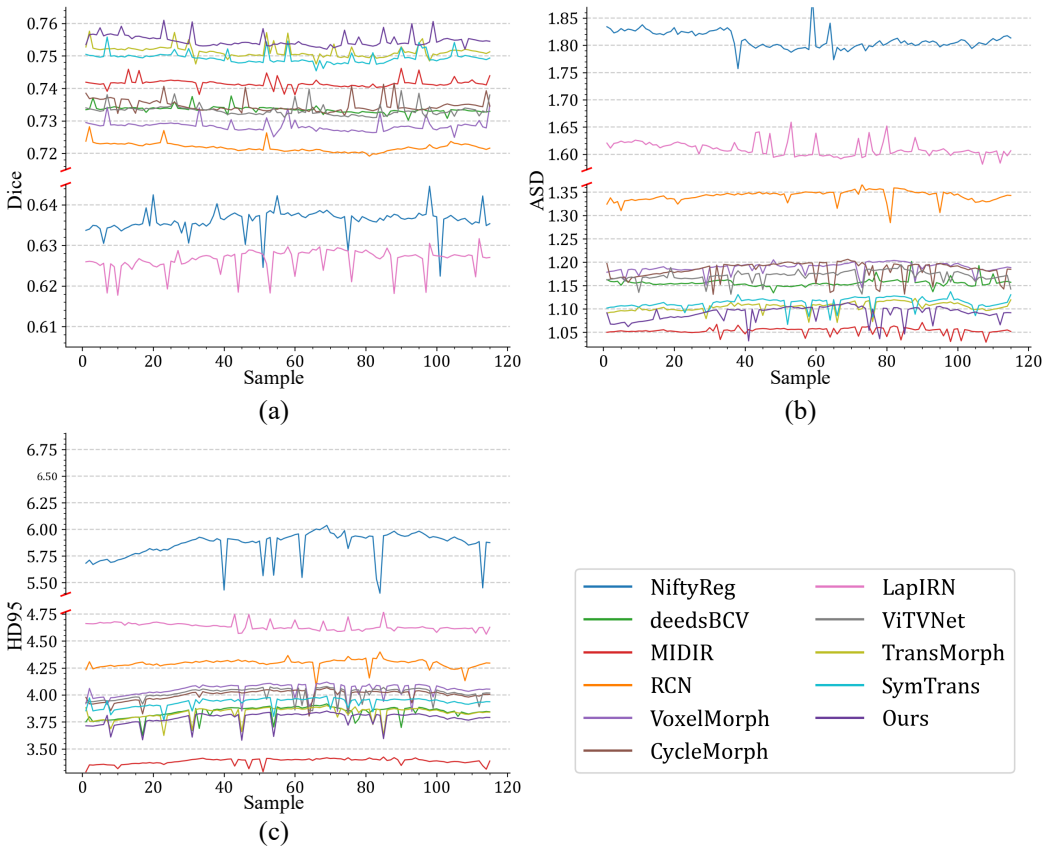


Figure 6: Curves of Dice (a), ASD (b), and HD95 (c) for all samples in test set obtained with different methods on the IXI dataset, with different colors indicating different methods. For better visualization, the y-axis of three curve plots is broken due to the large gap in metrics bettween the NiftyReg and others.

registration algorithms are more stable. Except for LapIRN, we noticed that almost all the transformer-based models were better than convolution-based models, demonstrating that the long-range dependencies in transformer is beneficial for promoting the image registration performance, especially for our proposed MD-SGT network, it achieved the second highest Dice (0.782) (a little lower than that of LapIRN (0.786)), as well the lowest ASD (0.571) and the second lowest HD95 (1.429), which are improved by 1.4%, 4.9% and 5.7% respectively by comparing the suboptimal transformer-based model (SymTrans). Moreover, the standard deviations of all the evaluation metrics obtained by the proposed MD-SGT network were the smallest among the transformer-based methods, although it did not outperform the traditional methods and LapIRN, its stability was greatly improved comparing with transformer-based methods.

4.3. Ablation experimental results

The registration results for all the ablation studies are shown in Figure 9 and the corresponding radar plots of the quantitative evaluation metrics are given in Figure 10. For better visualization, some regions with obvious differences between different ablation studies are zoomed-in in Figure 9, and in the radar plots, the negative ASD and HD95 distances, as well as the Dice score are normalized to the range of [0.6, 1], the larger the area enclosed by the triangle, the better the registration performance. As demonstrated in Figure 9(a), both MBConv and MD-GTB modules are beneficial for aligning the warped images with reference, especially using them simultaneously can better match the contours of the warped hippocampus and thalamus with those in the reference. Such improvements are also clearly demonstrated in the Figure 10(a), comparing with the base model (green triangle), introducing both MBConv (cyan triangle) and MD-GTB (blue triangle) modules can improve the Dice Score, ASD and HD95 distances. Specifically, introducing MBConv can decrease mean ASD/HD95 distances from 1.111/3.872 to 1.098/3.809, and increase the mean Dice score from 0.752 to 0.753, verifying that the stride convolution used in MBConv can overcome the problem of maxpooling in missing some information, therefore, it is able to extract useful features for promoting the registration performance; while introducing MD-GTB module can decrease the mean ASD/HD95 distances from 1.111/3.872 to 1.072/3.700 and increase the mean Dice score from 0.752 to 0.755, which indicates that MD-GTB is more useful than MBConv for registration. Combining the MBConv and MD-GTB modules (red triangle) can further promote the registration performance, the mean Dice score, ASD and

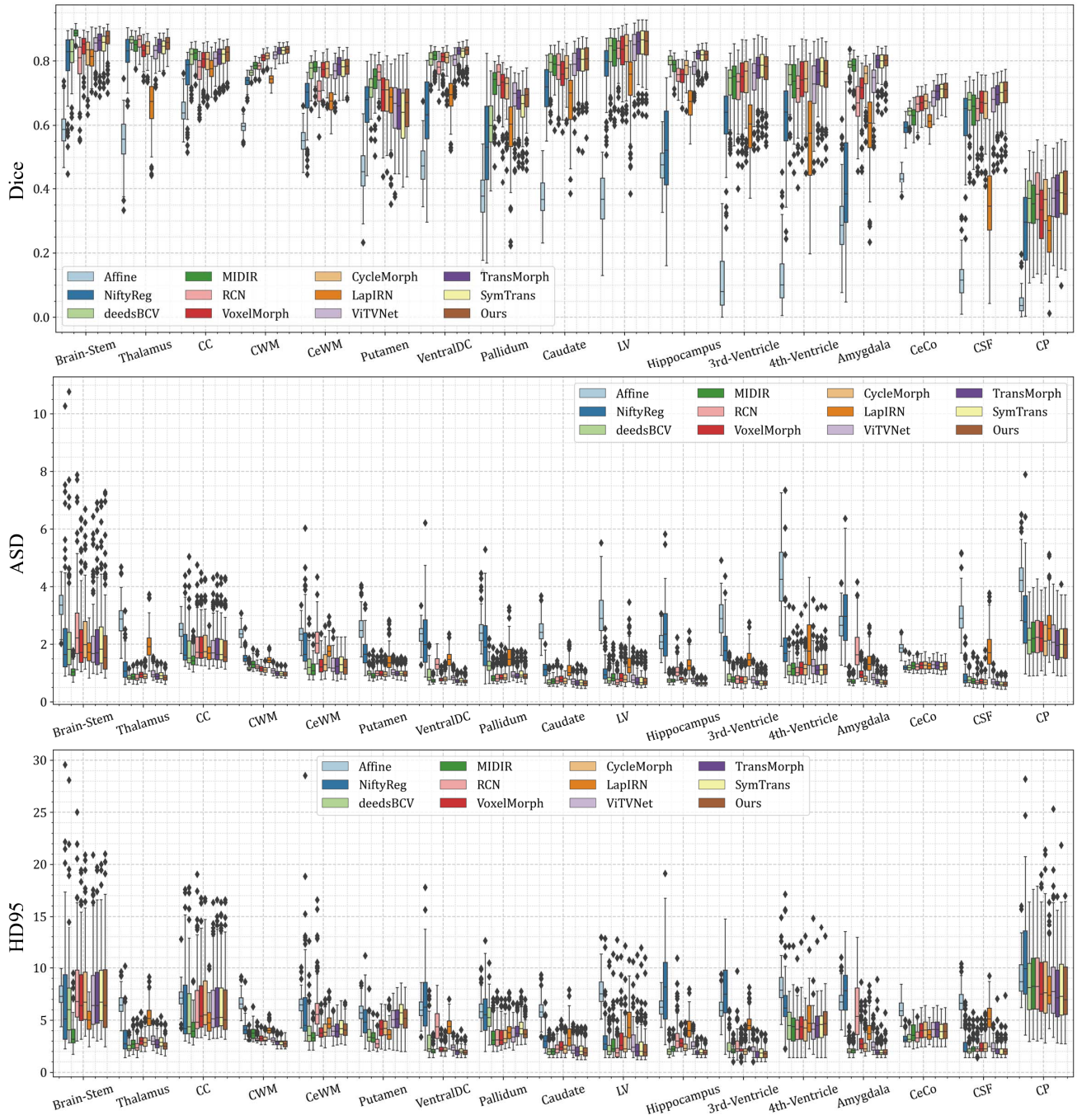


Figure 7: Boxplots of Dice scores (top), ASD (middle) and HD95 (bottom) distances for different brain areas using the proposed MD-SGT and the state-of-the-art registration methods.

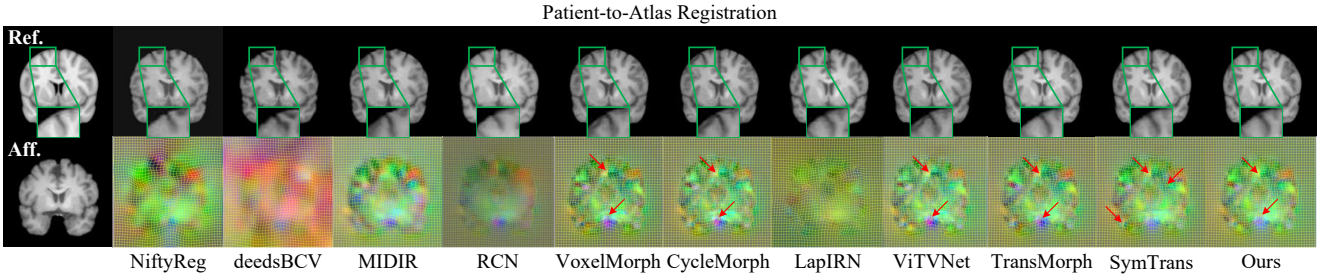


Figure 8: Qualitative comparison between the reference and the warped images obtained with different methods on the in-house dataset. The red arrows indicate the folding in the deformation field.

Table 5
Quantitative comparisons among different methods on patient-to-atlas registration.

Method	Patient-to-Atlas			
	Dice↑	ASD↓	HD95↓	% of $\lvert J_\phi \rvert \downarrow$
Affine	0.626±0.066	0.994±0.252	2.382±0.517	-
NiftyReg (Modat et al. (2010))	0.702±0.009	0.724±0.028	1.930±0.080	0.175±0.052
deedsBCV (Heinrich et al. (2013))	0.751±0.011	0.631±0.022	1.613±0.086	< 0.0001
MIDIR (Qiu et al. (2021))	0.748±0.043	0.660±0.112	1.677±0.314	< 0.0001
RCN (Zhao et al. (2019a))	0.740±0.030	0.665±0.070	1.702±0.190	0.220±0.010
VoxelMorph (Balakrishnan et al. (2018))	0.760±0.039	0.627±0.095	1.580±0.255	0.172±0.030
CycleMorph (Kim et al. (2021))	0.763±0.040	0.617±0.100	1.579±0.284	0.145±0.050
LapIRN (Mok and Chung (2020))	**0.786±0.010**	0.574±0.020	**1.325±0.080**	< 0.0001
ViTVNet (Chen et al. (2021))	0.763±0.045	0.621±0.120	1.581±0.334	0.158±0.027
TransMorph (Chen et al. (2022))	0.767±0.046	0.611±0.122	1.553±0.324	0.191±0.028
SymTrans (Ma et al. (2022))	0.771±0.037	0.599±0.093	1.511±0.284	0.239±0.026
MD-SGT (Ours)	0.782±0.024	**0.571±0.060**	1.429±0.174	0.149±0.020

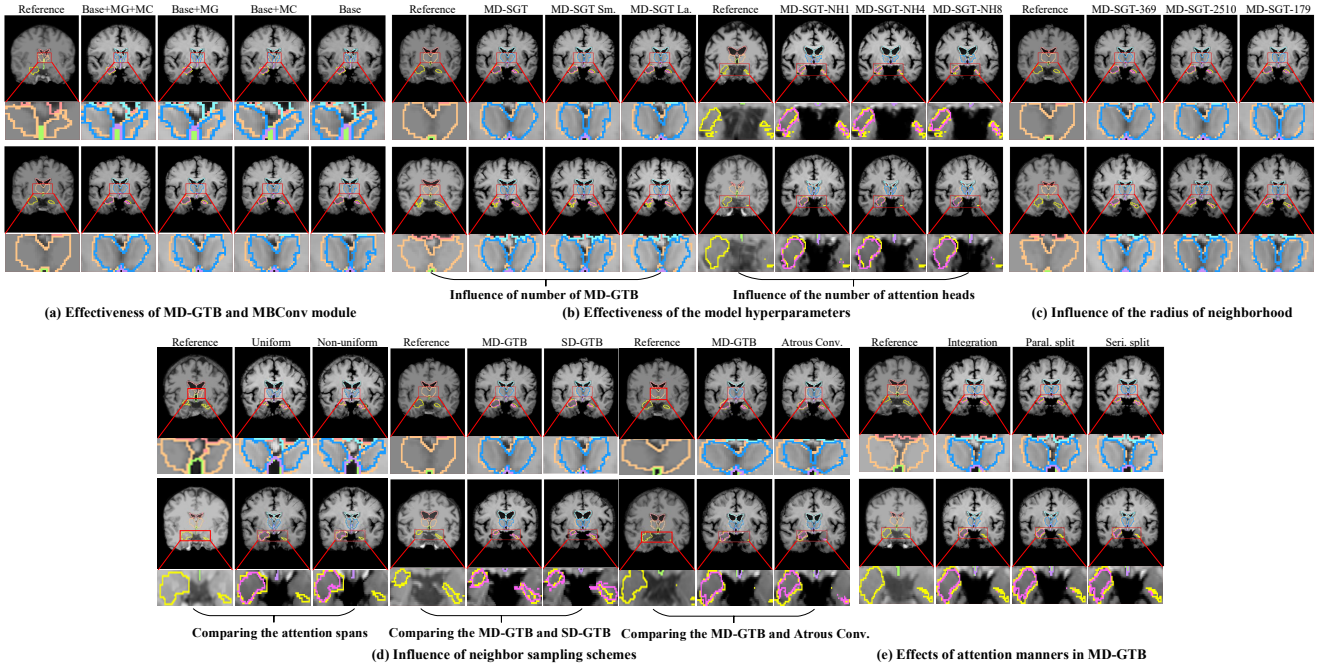
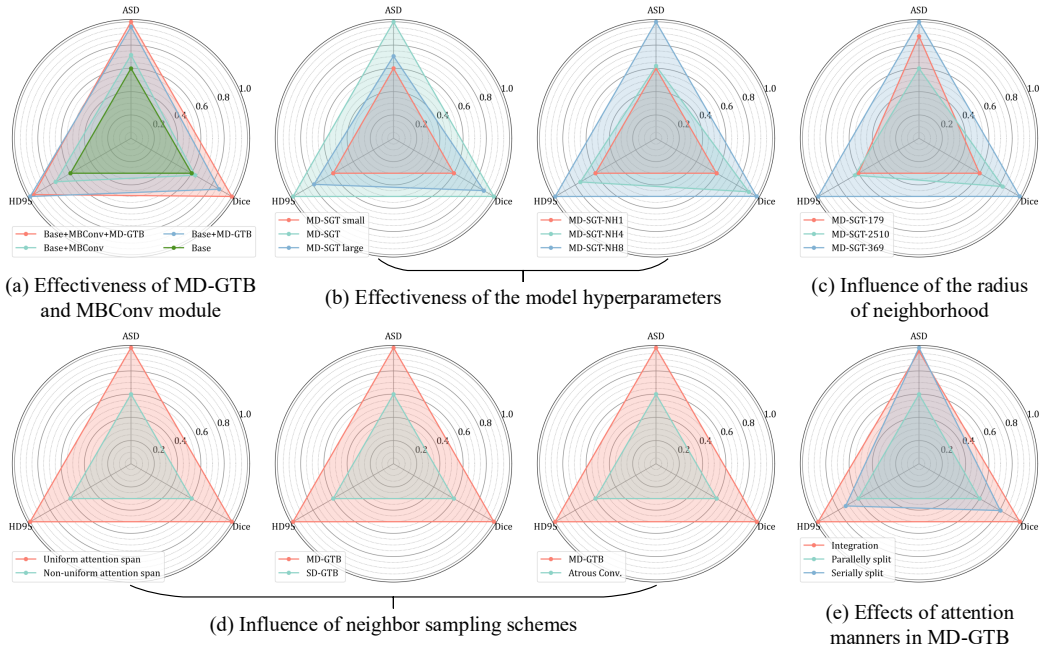


Figure 9: Qualitative comparison between the reference and warped images of ablation studies on the IXI dataset. Boundaries of several anatomical structures were overlaid on the warped images.

HD95 distances are improved by about 1%, 4%, and 4%, respectively, relative to the base model.

The number of MD-GTB modules, the feature embedding dimensions used in MD-GTB, and the number of attention heads are the main factors that influence the model size and accordingly affect the registration performance. As demonstrated Figure 10(b), we notice that increasing simultaneously the number of MD-GTB modules and the feature embedding dimensions cannot improve the registration performance (comparing MD-SGT large (blue triangle) and MD-SGT (green triangle)), with the mean Dice score decreasing a little bit from 0.757 to 0.756, and the ASD/HD95 distance increasing from 1.068/3.714 to 1.084/3.749, respectively. However, keeping the number of MD-GTB modules unchanged and increasing the feature embedding dimensions from 48 to 96 (comparing MD-SGT small (red triangle) and MD-SGT (green triangle)), the mean Dice score is increased by 0.5% (from 0.753 to 0.757), ASD and HD95 are decreased by 2.1 % (from 1.090 to 1.068) and 1.8% (from 3.781 to 3.714), respectively. Such findings can also be reflected by the qualitative registration results, as shown in Figure 9(b), comparing with MD-SGT small and MD-SGT large, MD-SGT achieves the best registration results, with the outline of the warped thalamus (blue) being closest to the ground-truth (orange outline) in the reference image.

As to the influence of attention heads, we fixed both the number of MD-GTB modules and feature embedding dimensions and then changed the number of attention heads, the registration results were shown in Figure 9(b) and the corresponding quantitative comparison was given in Figure 10(b). We noticed that, Dice score is almost uninfluenced by the the numbers of attention heads (changing from 0.756 to



(a) Effectiveness of MD-GTB and MBConv module

(b) Effectiveness of the model hyperparameters

(c) Influence of the radius of neighborhood

(d) Influence of neighbor sampling schemes

(e) Effects of attention manners in MD-GTB

Figure 10: Radar plots of ablation studies on the IXI dataset. Note that the negative ASD and HD95 distances, as well as the Dice score are normalized to the range of [0.6, 1]. The larger the area of the triangle, the better the registration performance.

0.757 when the number of attention heads varies from 1 to 8), but the ASD and HD95 distances improve gradually, with ASD decreasing from 1.081 to 1.054 and HD95 decreasing from 3.737 to 3.678. These findings indicate that increasing the attention heads is useful for improve the match degree of region boundaries and surfaces between the reference and warped images, as shown in the zoomed-in regions of Figure 9(b).

In this work, the MD-GTB module is the key element of the proposed registration network, in which the graph neighbor sampling schemes, the sampling radius, as well as the attention calculation manners are the dominant factors that affect the registration performance. As shown in Figure 9(c), we found that when the sampling radii at three scales are set as 3, 6 and 9 (MD-SGT-369) respectively, the model achieves the best performance, with the highest dice score (0.757) and the lowest ASD (1.068) and HD95 distances (3.714), meanwhile, the contours of hippocampus and thalamus in the warped image are closest to those in the reference image. As the sampling radii at three scales are changed to 1, 7, 9 (MD-SGT-179) and 2, 5 ,10 (MD-SGT-2510), the registration performance decreases quickly, as shown in Figure 10(c). In addition, from Figure 9(c) and Figure 10(c), we observed that there is no significant difference in performance between MD-SGT-179 and MD-SGT-2510, suggesting that the sampling radius is not the immediate cause that influences the performance, but the number of neighbor nodes determined by different radii affects the registration accuracy. In model MD-SGT-369, the number of neighbor nodes are 317, however, in model MD-SGT-179 and MD-SGT-2510, the number of neighbor nodes are 251 and 259 respectively. The more the neighboring nodes, the better the registration performance.

Keeping the number of neighboring nodes unchanged, Figure 9(d) shows the influence of different neighboring nodes sampling strategies, namely sampling neighbors with non-uniform and uniform attention spans, as well as with multi-dilations (MD-GTB) and single dilation (SD-GTB) at four scales, respectively. We found that the uniform attention span achieved the better registration results, with the outline of the warped thalamus (blue) and hippocampi (pink) being closest to the ground-truth (orange and yellow). Generally, using the uniform attention span, the mean Dice score increases from 0.753 to 0.756, and the ASD/HD95 distance decreases by 2.9% and 2%, respectively. When comparing the sampling strategies with multi-dilations and single dilation (the middle image in Figure 9(d)), we noticed that sampling the nodes from different spheres (multi-dilations) can effectively promote the registration performance, with some small region structures aligning well with the reference. This can also be intuitively revealed by the radar plots (Figure 10(d)), the Dice score, ASD and HD95 distances of MD-GTB are improved by 0.4%, 2.8% and 1.9% respectively comparing against with SD-GTB. Even though the idea of MD-GTB sampling strategy is similar to that of atrous convolution, its performance is much better than atrous convolution, especially in hippocampi and thalami regions (blue and pink regions in the right image of Figure 9(d)), with the Dice increased by 0.5%, and ASD and HD95 decreased by 2.1% and 1.6% respectively.

Besides the sampling radii at different scales and the sampling strategies, how to calculate the attention maps between the target node and the corresponding neighboring

nodes sampled from different FOVs also affects the registration results. As demonstrated in Figure 9(e) and Figure 10(e), calculating the attention maps with parallelly or serially splitting manners produced worse registration performance than with integration manner, with all the evaluation metrics decreased and the boundaries of some anatomical structure not well aligned with the reference.

5. Discussion

In this work, we proposed a novel deep learning model, MD-SGT, for unsupervised deformable medical image registration, in which the multi-dilation graph Transformers (MD-GTB) and convolutional blocks were combined enabling the model to benefit from the long-range learning ability of self-attention mechanism as well as the convolution inductive bias. Through the comparisons with the conventional and state-of-the-art learning-based methods on two datasets, we demonstrated that such hybrid network can deal with the problem of non-uniformed attention span occurred usually in transformer-based registration models, achieving an outstanding performance.

It is well known that the transformer-based models can learn more useful features due to their abilities of modeling long-range spatial dependencies and aggregating adaptively the context information, therefore, they may achieve better registration performance than convolution-based methods (Figure 5 and Table 4). To further validate that the long-range spatial dependence is indeed useful for deformable image registration, we plotted the effective receptive field maps of different registration methods obtained with the gradient back-propagation(Luo et al. (2016)) We observed that, MIDIR, RCN, VoxelMorph, and CycleMorph methods have the small receptive field size, while the LapIRN, ViTVNet, TransMorph, SymTrans and our method have the large receptive filed of almost the same size with the input image. By comparing their Dice scores, it can be seen that most of the methods with large receptive filed size outperforms those with small receptive field on both datasets. This can be explained that larger receptive filed can be aware of the boundary of the skull (just as shown in Figure 11), which may provide beneficial constraints on the deformation field. Even though the CNN-based method LapIRN and the transformer-based method ViTVNet can also achieve the large receptive field, their performances are worser on IXI dataset while better on in-house dataset. This illustrates that these two methods are sensitive to the image size or the texture information. This can be understood in terms of the nature of these registration methods. In LapIRN, it proposes a coarse-to-fine registration strategy by feeding the coarse velocity field into the registration framework of the next scale and adding the velocity fields at different scales to derive the final one. If the velocity field at the coarse scale is not accurate, the error will be propagated into the fine scales and therefore influences the results. In addition, the contributions of the velocity fields derived from different scales for the final velocity field should be different, adding them directly will also bring some bias. If the image size is small or the image texture is relatively simple, the error in velocity field at coarse scale is not significant, accordingly, the effect of error propagation is not obvious, that is why LapIRN performs better on in-house dataset with smaller image size. In ViTVNet, the transformer block is applied on the bottleneck layer, although it can reduce the computation complexity and enable the model has long-range spatial dependence, using the coarse attention coefficients calculated at the bottleneck layer to infer the dense predictions of velocity field is easy to generate errors. The larger the original image size or the more complex the image texture, the greater the difference between the attention coefficients calculated by the bottleneck layer and those calculated by the original image. Since in IXI dataset, the image size is much larger than that of in-house dataset, the influence of the bias in attention coefficients is accordingly more significant, this is why the ViTVNet performs worse on IXI dataset.

To deal with the above-mentioned problem of ViTVNet and to reduce also the computation complexity of the transformer applied on whole image, in TransMorph and SymTrans, the window-based attention is used. However, the attention span of the window-based attention is not uniform, which will limit the model to effectively extract the useful features for registration. As illustrated in Figure 3(a), in window-based attention, the target node in a given window can only aggregate the information from the other nodes in this window, which means that for the different target nodes in a window, the location distribution of their neighboring nodes are totally different. For instance, the neighbors of the red target node in Figure 3 (a) are mostly from its bottom right region, while the neighbors of the blue target node are mostly from its top left region. However, in our proposed MD-GTB, the neighbors of any target node are from its self-centered regions (Figure 3 (b)), allowing the model to update the target node information in a uniform manner. In addition, in MD-GTB, to enlarge the receptive filed size, the neighboring nodes are sampled from the several spheres with multiple dilation rates. Therefore, with the help of large receptive field size and uniform attention span, the performance of the proposed method is better than TransMorph and SymTrans (Figure 5 and Figure 8).

Although the multiple spherical neighborhood sampling strategy can promote the registration performance, its improvement is determined by the sampling radius for different dilation rates. From the ablation study (Figure 10(c)), we surprisingly found that, corresponding to the dilation rates of (1, 2, 3), our model achieved the best performance when the sampling radius are set as (3, 6, 9), rather than (1, 7, 9) and (2, 5, 10). That means, when sampling with low dilation rates (1 and 2), increasing the corresponding sampling radius can improve the registration performance. This can be intuitively explained, such sampling manner can sample more neighbors for each target node, accordingly, the target node can aggregate more useful information. When keeping the number of neighboring nodes unchanged but varying the sampling regions, which means sampling the same amount

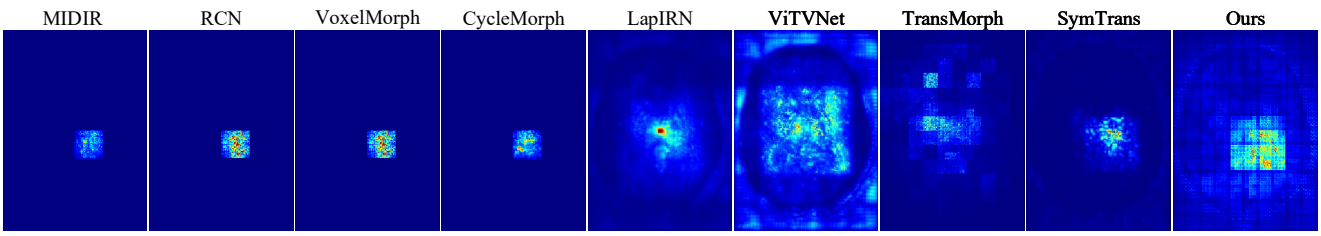


Figure 11: The effective receptive field maps of different registration methods obtained with salience map of a given voxel in bottleneck layer of each model.

of neighbors from a small region with dilation rate of 1 (SD-GTB) or from a bigger region with multiple dilation rates (MD-GTB), we found that MD-GTB performs better since it can capture long range dependence or more global context information.

It is well known that transformer is not the unique way to enlarge the receptive field. In the work of Jia et al. (2022), they proposed a large kernel UNet (LKU-Net) to implement the registration, they found that increasing the convolution kernel size can increase the receptive field and therefore be able to promote the registration performance, but it is difficult to control the kernel size to derive the appropriate receptive field size. If the kernel size is too small, the receptive field is not large enough to derive the accurate deformation field, while if the kernel size is too large, the receptive filed is much larger than the input image which may degrade the registration results. In contrast, there is no such problem in transformer-based architecture, the global operation in transformer enables us to see the whole image information, therefore it is not necessary to choose the appropriate kernel size to derive the right receptive field. Besides the large-kernel convolutions, using atrous convolution can also realize the long-range spatial dependence. We notice that the sampling strategy used in MD-GTB is like that of atrous convolution, but the performance of MD-GTB is much better (Figure 10(d)). This can be explained by their information aggregation manner. In MD-GTB, thanks to the attention mechanism, the information of each node can be adaptively updated according to the neighboring information, however, in the atrous convolution, the information of all the nodes (voxels) is updated in a fixed way (convolution kernels are shared everywhere), therefore, MD-GTB may capture more useful information than atrous convolution, accordingly, achieving the better performance.

Besides, we also take a deeper look into the effects of different attention manners used in MD-GTB on registration performance. As illustrated in Figure 4, we observed that in integration attention manner, the target node (red point) simultaneously grabs information from different FOVs, while for the other two manners, the target node aggregates information from different FOVs either in parallel or sequentially, resulting in the information of some neighboring nodes is utilized multiple times, such redundancy information deceases therefore the registration performance.

Even though the comparative experimental results have demonstrated the effectiveness of the proposed MD-GTB, how to set the hyper-parameters of MD-GTB is not trivial. As indicated by the ablation results (Figure 9), we noticed that some settings of MD-GTB significantly influenced the registration performance. First of all, the number of MD-GTB modules and the feature embedding dimensions used in MD-GTB are the determinative factors in model size which in turn affects the registration performance directly. Generally, the larger model size produces better performance. However, in our work, increasing the model size at the beginning, the registration performance improves (changing from MD-SGT-small to MD-SGT), but as increasing the model size continually, the registration performance decreases instead (changing from MD-SGT to MD-SGT-large). This phenomenon can be explained with the under-parameterized regime proposed by Nakkiran et al. (2021), in which the authors argued that the variation of model performance with the model size follows the U-like behavior, that means increasing model complexity will increase performance first, and then decrease it when the model complexity passes a certain threshold, that is why the performance of MD-SGT-larger is not better than MD-SGT as expected. In addition, increasing the number of attention heads in MD-GTB will also increase the model size, but in this work, increasing the attention heads from 1 to 8 does not make model complexity pass the threshold mentioned above. Meanwhile, as indicated by Vaswani et al. (2017), multi-head attentions enable the model extracting rich information from different representation subspaces at different positions, therefore increasing the number of attention heads can improve the registration performance.

All the comparison and ablation results demonstrated that the proposed MD-SGT is more advantageous for deformable registration than the existing CNN-based and Transformer-based models. However, there are still several limitations need to be addressed in the future. First, in the proposed MD-GTB module, only the distance between the target node and potential neighboring nodes was considered when constructing the local graph, although it is simple and efficient, it overlooks the features of each node. Therefore, using some node-feature guided similarity measures to construct the local graph may be our future work. Second, considering that the distance between the nodes may also influence the information aggregation, thus taking account the distance into the self-attention mechanism of a multi-dilation graph is also our interest. Third, the neighboring

nodes used in this work were sampled uniformly from different spherical regions, how to adaptively sampling the neighbors to further improve the registration performance also need to be considered. Finally, although the proposed model solved the problem of non-uniform attention span of the window-attention based methods, as well as the issues of coarse attention map in transformer used on bottleneck layer, the computation complexity and the model parameters are much larger, as illustrated in Figure 12. In MD-GTB, graph attention was calculated in a node-by-node manner to achieve the uniform attention span. That means each node has its own local graph, and the number of the local graphs is determined by the number of voxels in each feature map, such huge graph numbers and the attention calculations for each graph increase undoubtedly the computation time and learning parameters. How to retain the superiority of such local graph attention with uniform attention span, as well as to reduce the computation complexity of the MD-SGT will be of interest.

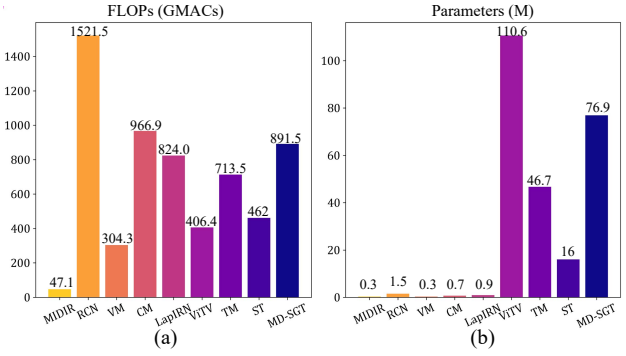


Figure 12: The comparisons among different learning-based registration models in terms of computation complexity on IXI dataset and number of learning parameters. FLOPs or GMACS, giga multiply–accumulate operations was used as the metric to reflect the computation complexity. TM:TransMorph, CM:CycleMorph, VM: VoxelMorph, ViTV: ViTVNet, and ST: symTrans.

6. Conclusion

To deal with the problems of the limited long-range spatial dependence and non-uniform attention span in the existing registration models, we proposed a multi-dilation spherical graph transformer (MD-SGT) for unsupervised deformable medical image registration, in which the differences between the reference and the template images at various scales were fully extracted by combining the convolution inductive bias and long-range uniform attention span of graph transformer in the encoder, and then the final deformation filed was estimated from such feature differences using a decoder. By comparing the proposed model with the state-of-the-art methods for two registration tasks, atlas-to-patient and patient-to-atlas registrations, on different datasets, we demonstrated the superiority and effectiveness of MD-SGT, the DICE, ASD and HD95 were improved at least by 0.5%, 2.2% and 1.1%, respectively. Comparing with the SOTA methods, its stability is the best.

Acknowledgments

This work was partially funded by the National Natural Science Foundations of China (Grant No.62161004), Guizhou Provincial Science and Technology Projects (QianKeHe ZK [2021] Key 002), and Guizhou Provincial Science and Technology Projects (QianKeHe ZK [2022] 046).

References

Arsigny, V., Commowick, O., Pennec, X., Ayache, N., 2006. A log-euclidean framework for statistics on diffeomorphisms, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 924–931.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis 12, 26–41.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450 .

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. An unsupervised learning model for deformable medical image registration, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9252–9260.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging 38, 1788–1800.

Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision 61, 139–157.

Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., Shen, D., 2017. Deformable image registration based on similarity-steered cnn regression, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 300–308.

Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. Transmorph: Transformer for unsupervised medical image registration. Medical Image Analysis 82, 102615.

Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y., 2021. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468 .

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018. Unsupervised learning for fast probabilistic diffeomorphic registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 729–738.

Devalla, S.K., Renukanand, P.K., Sreedhar, B.K., Subramanian, G., Zhang, L., Perera, S., Mari, J.M., Chin, K.S., Tun, T.A., Strouthidis, N.G., et al., 2018. Drunet: a dilated-residual u-net deep learning network to segment optic nerve head tissues in optical coherence tomography images. Biomedical optics express 9, 3244–3265.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Fischl, B., 2012. Freesurfer. Neuroimage 62, 774–781.

Hassani, A., Walton, S., Li, J., Li, S., Shi, H., 2022. Neighborhood attention transformer. arXiv preprint arXiv:2204.07143 .

Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A., 2013. Mrf-based deformable registration and ventilation estimation of lung ct. IEEE transactions on medical imaging 32, 1239–1248.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. Advances in neural information processing systems 28.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841.

Jenkinson, M., Pechaud, M., Smith, S., et al., 2005. Bet2: Mr-based estimation of brain, skull and scalp surfaces, in: Eleventh annual meeting of the organization for human brain mapping, Toronto.. p. 167.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical image analysis 5, 143–156.

Jia, X., Bartlett, J., Zhang, T., Lu, W., Qiu, Z., Duan, J., 2022. U-net vs transformer: Is u-net outdated in medical image registration?, in: Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings, Springer. pp. 151–160.

Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C., 2021. Cyclemorph: cycle consistent unsupervised deformable image registration. Medical Image Analysis 71, 102036.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 .

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2009. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging 29, 196–205.

Kuang, D., Schmah, T., 2019. Faim–a convnet method for unsupervised 3d medical image registration, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 646–654.

Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R., 2021. Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511 .

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Luo, W., Li, Y., Urtasun, R., Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems 29.

Luo, Y., Cao, W., He, Z., Zou, W., He, Z., 2021. Deformable adversarial registration network with multiple loss constraints. Computerized Medical Imaging and Graphics 91, 101931.

Ma, M., Xu, Y., Song, L., Liu, G., 2022. Symmetric transformer-based network for unsupervised image registration. Knowledge-Based Systems 257, 109959.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), IEEE. pp. 565–571.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine 98, 278–284.

Mok, T.C., Chung, A.C., 2020. Large deformation diffeomorphic image registration with laplacian pyramid networks, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23, Springer. pp. 211–221.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., Sutskever, I., 2021. Deep double descent: Where bigger models and more data hurt. Journal of Statistical Mechanics: Theory and Experiment 2021, 124003.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.

Qiu, H., Qin, C., Schuh, A., Hammernik, K., Rueckert, D., 2021. Learning diffeomorphic and modality-invariant registration using b-splines, in: Medical Imaging with Deep Learning.

Smith, S.M., 2002. Fast robust automated brain extraction. Human brain mapping 17, 143–155.

Sokooti, H., Vos, B.d., Berendsen, F., Lelieveldt, B.P., Išgum, I., Staring, M., 2017. Nonrigid image registration using multi-scale 3d convolutional neural networks, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 232–239.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y., 2022. Maxvit: Multi-axis vision transformer. arXiv preprint arXiv:2204.01697 .

Uzunova, H., Wilms, M., Handels, H., Ehrhardt, J., 2017. Training cnns for image registration from few samples with model-based data augmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 223–231.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., 2017. Graph attention networks. arXiv preprint arXiv:1710.10903 .

Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage 45, S61–S72.

Vos, B.D.d., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I., 2017. 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. Springer, pp. 204–212.

Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., et al., 2019. Deep graph library: A graph-centric, highly-performant package for graph neural networks. arXiv preprint arXiv:1909.01315 .

Yang, X., Kwitt, R., Niethammer, M., 2016. Fast predictive image registration, in: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 48–57.

Zhang, L., Ning, G., Zhou, L., Liao, H., 2023. Symmetric pyramid network for medical image inverse consistent diffeomorphic registration. Computerized Medical Imaging and Graphics , 102184.

Zhao, S., Dong, Y., Chang, E.I., Xu, Y., et al., 2019a. Recursive cascaded networks for unsupervised medical image registration, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10600–10610.

Zhao, S., Lau, T., Luo, J., Eric, I., Chang, C., Xu, Y., 2019b. Unsupervised 3d end-to-end medical image registration with volume tweening network. IEEE journal of biomedical and health informatics 24, 1394–1404.