A TRANSFORMER-BASED NETWORK FOR DEFORMABLE MEDICAL IMAGE REGISTRATION

Yibo Wang, Wen Qian, and Xuming Zhang

College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

ABSTRACT

Deformable medical image registration plays an important role in clinical diagnosis and treatment. Recently, the deep learning (DL) based image registration methods have been widely investigated and showed excellent performance in computational speed. However, these methods cannot provide enough registration accuracy because of insufficient ability in representing both the global and local features of the moving and fixed images. To address this issue, this paper has proposed the transformer based image registration method. This method uses the distinctive transformer to extract the global and local image features for generating the deformation fields, based on which the registered image is produced in an unsupervised way. Our method can improve the registration accuracy effectively by means of self-attention mechanism and bi-level information flow. Experimental results on such brain MR image datasets as LPBA40 and OASIS-1 demonstrate that compared with several traditional and DL based registration methods, our method provides higher registration accuracy in terms of dice values.

***Index Terms*—** Image Registration, Deep Learning, Transformer, Registration Accuracy

1. INTRODUCTION

Image registration is one of the fundamental and challenging tasks in medical image processing and analysis. Its goal is to find the correspondence between the moving and fixed images to facilitate such tasks as disease diagnosis and surgical navigation. Up to now, the various image registration methods have been proposed. For the traditional registration methods [1–3], the similarity metric is firstly constructed between the fixed and moving images. Then, the objective function based on the constructed metric is optimized to produce the registered image. These methods are time-consuming because of the complicated iterative optimization.

To improve image registration efficiency, the deep learning (DL) based registration methods have been presented. Given numerous moving and fixed images, the deep neural networks can be trained to generate the registered image efficiently. Depending on how the networks are trained, these methods can be categorized into the supervised learning and unsupervised learning ones. In the supervised approaches, the ground-truth deformation fields or anatomical landmarks are needed [4–7]. Sokooti et al. [4] have proposed a convolution neuron network (CNN) to directly estimate the displacement vector field (DVF) using the artificially generated DVFs. Cao et al. [7] have developed a deformable inter-modality image registration method which estimates the deformation fields using the deep neural network supervised by intra-modality similarity. The registration performance of these methods greatly depends on the ground-truths which are generally difficult to acquire in clinical scenarios. As for the unsupervised learning based methods [8–11], they need no the ground truth of the deformation field. Balakrishnan et al. [8] have proposed a 3D medical image registration model, voxelmorph, which reconstructs the registered result using a CNN with a spatial transform layer. Zhao et al. [9] have designed a volume tweening network (VTN) including the cascaded subnetworks to improve the registration performance recursively. Kim et al. [11] have presented a cycle-consistent deformable image registration method called cyclemorph, which can enhance the registration performance by introducing the cycle consistency into the network loss.

Although the existing DL based registration approaches can provide higher computational efficiency than the traditional ones, they cannot capture the long-range dependence in the moving and the fixed image effectively because of the adoption of such networks as the CNN which has the limited ability of extracting the global image features. Therefore, the registration accuracy of these DL based methods is influenced disadvantageously especially when the large deformation is involved between the fixed and moving images. Recently, the transformer has become an important network in the fields of natural language processing and computer vision because it can explore the long-range dependence based on the self-attention mechanism. Distinctively, the transformer can extract the global image features effectively, thus it has been applied to such tasks as image classification [12] and image denoising [13]. To overcome the disadvantages of existing CNN based registration methods, we have presented a novel deformable medical image registration network called

This work was sponsored by CAAI-Huawei MindSpore Open Fund and the National Natural Science Foundation of China under Grant 61871440.

Transformer-UNet (TUNet). This network introduces the vision transformer (VIT) [12] into the framework of UNet [14] to extract the global and local features from the moving and fixed images, thereby generating the deformation field effectively. Besides, the skip connections are established in the bi-level layers to guarantee the correct information flow between the rough features and the fine features.

Experiments have been done on LPBA40 and OASIS-1 to test the performance of our method. The qualitative and quantitative evaluations demonstrate that the proposed method is provided with higher registration accuracy than the compared traditional and DL based registration methods.

The paper is organized as follows. Section 2 describes our method. Section 3 presents the experimental results of our method on two datasets. Conclusion is given in Section 4.

2. METHOD

The framework of the TUNet is shown in Fig. 1. Here, a moving image M and a fixed image F are input into the TUNet. The deformation field ϕ is computed based on the parameters learned in the different network layers. By means of the spatial transform layer, M is deformed to produce the registered image R. The TUNet is trained using the loss defined by the dissimilarity between F and M and the smoothness constraint of ϕ to produce the registered result in an unsupervised way.

2.1. Architecture of the Transformer-UNet

Our Transformer-UNet is built on the encoder-decoder architecture of the UNet [14], but improves the latter by introducing the bi-level connection and an unique Transformer block. As shown in Fig. 1, the proposed Transformer-UNet uses a single input formed by concatenating M and F in the dimension channel. In the encoder, the two Conv layers are used to extract image features, where each block is composed of a convolutional module followed by the Rectified Linear Unit (ReLU). The kernel size and stride in the convolutional module will be set to $3\times3\times3$ and 1, respectively. The Max-pooling layer, Conv layers and Transformer blocks are combined to produce image features at different levels. In some layers in the encoder, the concatenation layer is additionally introduced to concatenate the features produced by the Transformer block at the previous layer and those resulting from the pooling layer. In the decoder, the Conv layers, Transformer blocks and Up-sampling layer are combined to store the spatial resolution of the feature maps at different levels. The features at the same levels produced by the encoder and decoder will be concatenated by the concatenation layers. At the end of decoder, the concatenated features will be processed by the two Conv layers to output the final features. Note that the stride of Max-pooling layer and Up-sampling layer is set to 2, and thus the encoder reduces the spatial resolution of input volumes by a factor of 8 in total and the decoder restores the features to the original size. As the key component of our method, the Transformer block is distributed at different layers in the encoder and decoder. It receives the convolutional features and outputs two different feature maps.

2.2. Transformer Block

Inspired by the VIT [12], we will build the Transfomer block shown in Fig. 2. Compared with the VIT, this block retains the multi-head self-attention mechanism, which is necessary for improving network's awareness for the global information. Meanwhile, we have made some modifications on the VIT to produce the distinctive Transfomer block. Firstly, the redundant position embedding is removed after the patch embedding. Secondly, the convolution module is used to directly compute the weight matrix instead of the original patch embedding and the linear mapping, which will reduce the computational complexity and meet the need of 3D image registration better. Finally, an additional output path has been designed. By using the convolution module with a stride of 2 or the deconvolution module, we have established a bridge for information flow in the bi-level features.

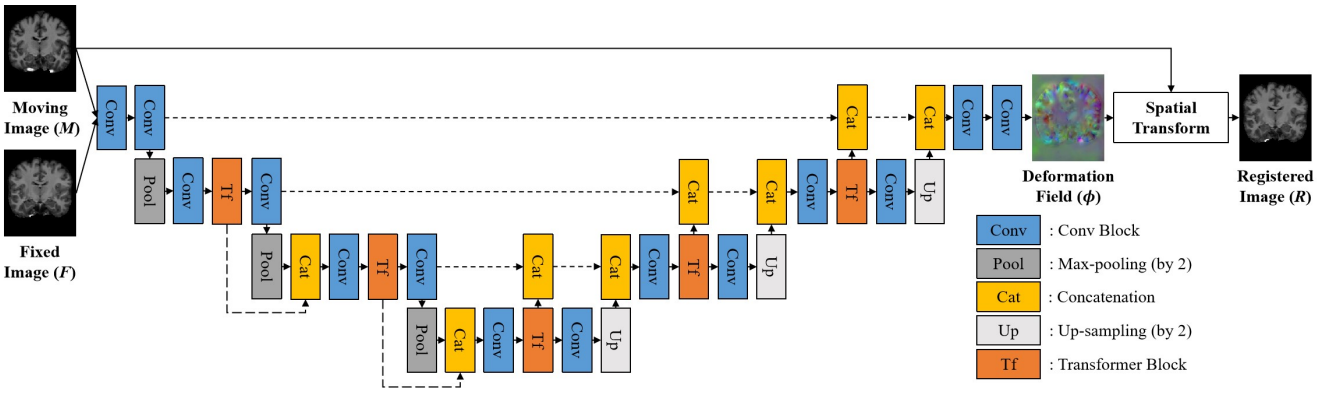

Fig. 1. The overall framework of the proposed method, Transformer-UNet, for deformable medical image registration. Here, the short and long dashed lines denote the skip connection and the bi-level connection, respectively.

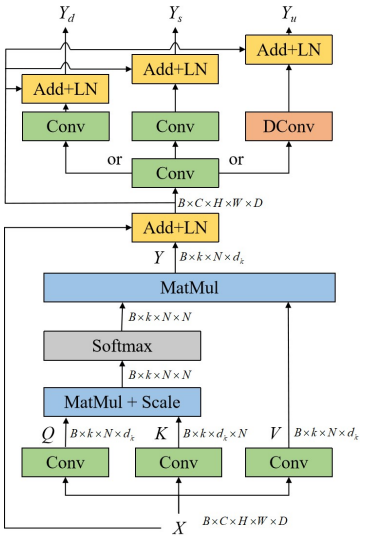

Fig. 2. The structure of the proposed Transformer block.

As shown in Fig. 2, the input feature X is processed by the three different convolution modules to generate the query matrix Q, the key matrix K and the value matrix V as:

$$Q = W^Q \cdot X; K = W^K \cdot X; V = W^V \cdot X, \quad (1)$$

The matrices Q, K and V are reshaped into a sequence of flattened 3D patches: Q, K, and $V \in \mathbb{R}^{B\times N\times(P^3\cdot C)}$, where B is the mini-batch, C is the number of channels, (P, P, P) is the resolution of each volume patch, and $N = HWD/P^3$ is the resultant number of patches with (H, W, D) denoting the resolution of the input feature. By splitting the heads from the embed channels and swapping the order of axis, we will change them into a 4D vector: Q, $V \in \mathbb{R}^{B\times k\times N\times d_k}$, $K \in \mathbb{R}^{B\times k\times d_k\times N}$, where k is the number of heads and $d_k = (P^3 \cdot C)/k$ is the number of the embedding channels per head. The matrix multiplication, scaling and softmax operations will be implemented for Q, K and V to produce the output Y.

$$Y = softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad (2)$$

The output Y will be added to the input X and then is processed by the layer-norm (LN) operation. To promote the information interaction between two different levels of features, we will add the convolution module with a stride of 2 or the deconvolution (DConv) module at the end of block. Similarly, the LN will be applied again. In this way, our Transformer block will generate two output feature maps. One called Y_s has the same size as the input feature map while another Y_d or Y_u has half or twice the size of the input feature map.

2.3. Spatial Transform

For the spatial transform, we will choose the 3D transformation function with the bilinear interpolation defined as:

$$M \circ \phi = \sum_{q\in G(\phi(p))} M(q) \prod_{d\in\{x,y,z\}} (1 - |\phi_d(p) - q_d|), \quad (3)$$

where p is a voxel, $G(\phi(p))$ means the 8-neighbors of $\phi(p)$ and $\circ$ is the spatial transform function.

2.4. Loss Function

The loss function of the TUNet includes a dissimilarity term related to the local cross correlation (CC) and a smoothness regularization term of ϕ, and it is defined as:

$$\mathcal{L}(M, F, \phi) = -CC(M \circ \phi, F) + \lambda \sum_{p\in\Omega} \|\nabla\phi(p)\|, \quad (4)$$

where λ is a hyper-parameter and Ω denotes the 3D volume and $CC(A, B)$ is computed as:

$$CC(A, B) = \sum_{v\in\Omega} \frac{(\sum_{v_i}(A(v_i) - \bar{A}(v))(B(v_i) - \bar{B}(v)))^2}{\sum_{v_i}(A(v_i) - \bar{A}(v))^2 \sum_{v_i}(B(v_i) - \bar{B}(v))^2}, \quad (5)$$

where v_i is chosen as the $9 \times 9 \times 9$ patch, $\bar{A}(v)$ and $\bar{B}(v)$ mean the local mean of $A(v)$ and $B(v)$, respectively.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1. Experimental Settings

Datasets We have chosen LPBA40 [15] and OASIS-1 [16] for experiments. The LPBA40 contains 40 T1-weighted brain MR images, where 56 anatomical areas are segmented from each image. The OASIS-1 contains 414 T1-weighted brain MR images, where each image includes 35 segmented cortical regions. Here, all scans are sampled to a 256×256×256 grid with 1mm isotropic voxel. The affine spatial normalization and brain extraction are carried out using FreeSurfer [17]. The images are further cropped into 192×160×192. For the LPBA40, we have trained our model on 25 subjects, validated it on 5 subjects and tested it on 5 subjects. For the OASIS-1, we have used 324 subjects for model training, 42 subjects for validation and 40 subjects for testing.

Implementation Details We will focus on atlas-based registration, in which a fixed volume is chosen as atlas and each volume in the dataset is registered to it. Here, because of high memory cost in the training stage, we will extract patches of size 128×128×64 from a whole volume, and set the corresponding batch size according to the GPU memory usage. To avoid over-fitting, the random rotation is implemented on each training volume pair to realize data augmentation. We set hyper-parameter λ to 0.1 and adopt Adam optimization with a learning rate of 1e-4. Our model is trained for 30 epochs on a single NVIDIA RTX 2080Ti GPU.

Compared Methods In order to verify the superiority of the TUNet, we will compared it with several popular image registration methods including SyN [2] from Advanced Normalization Tools, VoxelMorph (VM) [8], VTN [9] and CycleMorph (CM) [11]. As regards the VoxelMorph, we will choose VoxelMorph-1 [8] as the baseline network with the same parameters to our method for the fair comparison.

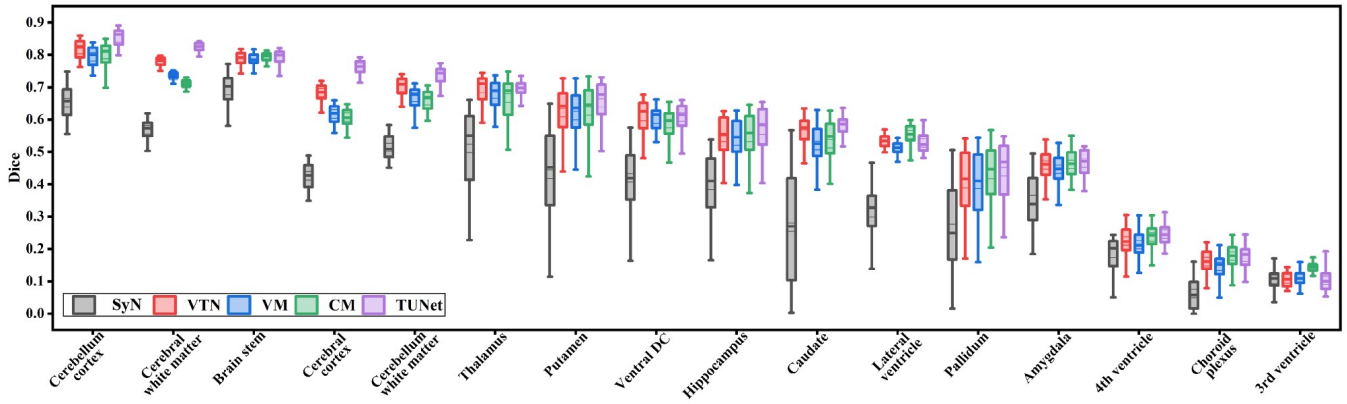

Fig. 3. Boxplots of Dice values for the SyN, VTN, VM, CM and TUNet performed on the anatomical structures.

Evaluation Metrics The registration performance is evaluated by Dice [18], which is defined as the overlap rate between the segmented results of registered and fixed images.

$$Dice(R, F) = 2 \cdot \frac{|R \cap F|}{|R| + |F|} \tag{6}$$

3.2. Qualitative Evaluation

Fig. 4 shows the registered results of the various methods operating on two chosen MR images from the OASIS-1 and LPBA40 datasets. The observation from Fig. 4 shows that the other compared methods cannot restore the deformation of some internal brain structures and brain contour effectively. By comparison, the proposed TUNet can provide the most similar registered results to the fixed image in terms of the different brain tissues among all evaluated methods. Especially, the proposed method can preserve some fine image details better than the other methods, which can be verified with the details in the bright yellow boxes. The superiority of the TUNet to other methods lies in its outperforming ability to extract both the global and local features, which indeed facilitates estimating the deformation fields more effectively.

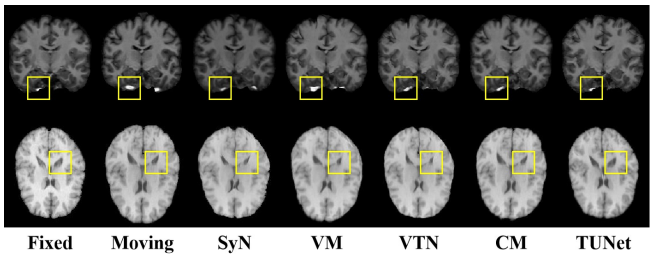

Fig. 4. The registered results of the various methods on OASIS-1 (the first row) and LPBA40 (the second row) .

3.3. Quantitative Evaluation

Fig. 3 visualizes the Dice values for 16 evaluated anatomical structures across all test samples in OASIS-1. For better visualization purpose, we combine the same structures from the left and right hemisphere together. It can be seen that the TUNet achieves higher Dice values than other methods in most of structures. Although the VTN and the CM provide the highest Dice values for Thalamus and Ventral DC as well as for Lateral ventricle and 3rd ventricle, respectively, our method can still provide the competitive Dice values for these structures. It should be especially emphasized that on such structures as Cerebellum cortex, Cerebral white matter, Cerebral cortex, Putamen and Hippocampus, our method significantly outperforms other methods in terms of Dice values.

Table 1 lists the Dice values of all evaluated methods on the four regions of images in the LPBA40. It can be seen from Table 1 that the TUNet outperforms the other methods in all regions. For example, the proposed method provides the improved Dice in Brain stem by 0.026 over the SyN and that in Hippocampus by about 5.6 % over the VM.

Table 1. Dice values of all evaluated methods on LPBA40.

Methods	Brain stem	Parietal	Hippocampus	Putamen
SyN	0.772	0.501	0.506	0.501
VM	0.781	0.554	0.518	0.544
VTN	0.791	0.582	0.516	0.547
CM	0.787	0.565	0.519	0.558
TUNet	**0.798**	**0.606**	**0.547**	**0.574**

4. CONCLUSION

In this paper, we have presented a Transformer-UNet based unsupervised deformable medical image registration method. The framework is built on the Transformer model which is introduced into the UNet. By means of the distinctive Transformer-UNet, the global and local features can be extracted from the moving and fixed images effectively, thereby ensuring good registration performance. The experimental results show that our method outperforms several traditional and deep learning based registration methods in terms of visual evaluations and such quantitative metrics as Dice. Future research will be focused on the extension of our method to multi-modal medical image registration.

5. REFERENCES

[1] John Ashburner, "A fast diffeomorphic image registration algorithm," *Neuroimage*, vol. 38, no. 1, pp. 95–113, 2007.

[2] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[3] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[4] Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring, "Nonrigid image registration using multi-scale 3d convolutional neural networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2017, pp. 232–239.

[5] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer, "Quicksilver: Fast predictive image registration–a deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.

[6] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec, "Svf-net: Learning deformable image registration using shape matching," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2017, pp. 266–274.

[7] Xiaohuan Cao, Jianhuan Yang, Li Wang, Zhong Xue, Qian Wang, and Dinggang Shen, "Deep learning based inter-modality image registration supervised by intra-modality similarity," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2018, pp. 55–63.

[8] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, Adrian V Dalca, and John Guttag, "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 9252–9260.

[9] Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al., "Recursive cascaded networks for unsupervised medical image registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2019, pp. 10600–10610.

[10] Yang Lei, Yabo Fu, Tonghe Wang, Yingzi Liu, Pretesh Patel, Walter J Curran, Tian Liu, and Xiaofeng Yang, "4d-ct deformable image registration using multiscale unsupervised deep learning," *Physics in Medicine & Biology*, vol. 65, no. 8, pp. 085003, 2020.

[11] Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye, "Cyclemorph: Cycle consistent unsupervised deformable image registration," *Medical Image Analysis*, vol. 71, pp. 102036, 2021.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Achleshwar Luthra, Harsh Sulakhe, Tanish Mittal, Abhishek Iyer, and Santosh Yadav, "Eformer: Edge enhancement based transformer for medical image denoising," *arXiv preprint arXiv:2109.08044*, 2021.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.

[15] David W Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L Narr, Russell A Poldrack, Robert M Bilder, and Arthur W Toga, "Construction of a 3d probabilistic atlas of human cortical structures," *Neuroimage*, vol. 39, no. 3, pp. 1064–1080, 2008.

[16] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[17] Bruce Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

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