# Learning Dual Transformer Network for Diffeomorphic Registration

Yungeng Zhang, Yuru Pei(✉), and Hongbin Zha

Department of Machine Intelligence, Key Laboratory of Machine Perception (MOE), Peking University, Beijing, China
peiyuru@cis.pku.edu.cn

**Abstract.** Diffeomorphic registration is widely used in medical image processing with the invertible and one-to-one mapping between images. Recent progress has been made to diffeomorphic registration by utilizing a convolutional neural network for efficient and end-to-end inference of registration fields from an image pair. However, existing deep learning-based registration models neglect to employ attention mechanisms to handle the long-range cross-image relevance in embedding learning, limiting such approaches to identify the semantically meaningful correspondence of anatomical structures. In this paper, we propose a novel dual transformer network (DTN) for diffeomorphic registration, consisting of a learnable volumetric embedding module, a dual cross-image relevance learning module for feature enhancement, and a registration field inference module. The self-attention mechanisms of DTN explicitly model both the inter- and intra-image relevances in the embedding from both the separate and concatenated volumetric images, facilitating semantical correspondence of anatomical structures in diffeomorphic registration. Extensive quantitative and qualitative evaluations demonstrate that the DTN performs favorably against state-of-the-art methods.

**Keywords:** Dual transformer · Diffeomorphic registration · Relevance learning

## 1 Introduction

Deformable registration is a fundamental and challenging problem in medical image processing, with the goal to find dense per-voxel displacement and establish alignment between a pair of images. The study of deformable registration has a variety of potential applications, especially in the analysis of multi-modal images captured from different subjects or in a longitudinal treatment with structure variations due to treatments and growths. The deformable registration produces the nonlinear voxel-wise mapping between images, facilitating the atlas-based annotation, statistical shape analysis, and shape comparison of anatomical structures. In order to accomplish the task of deformable registration effectively,


M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12904, pp. 129–138, 2021.
https://doi.org/10.1007/978-3-030-87202-1_13

we need to infer the semantic correspondence of fine-grained structures. The volumetric images vary in shapes, scales, and poses, so that it is a challenging issue to identify the real matching anatomical structures.

Traditional deformable registration is known to be computationally expensive due to iterative optimization of large-scale parameters [21]. Recently, the state-of-the-art convolutional neural network (CNN)-based methods has been proposed to address the deformable registration [4,9,12,18,26,27]. The CNN performs the end-to-end inference of the displacement or velocity fields from a pair of images, using regularization, such as the smoothness and the Jacobian determinant [14,18,26], for the invertible and the diffeomorphic transformations. Moreover, the symmetric registration infers a pair of diffeomorphic maps regarding the middle of the geodesic path [18]. However, these methods conduct a straightforward inference from the CNN-based low-level local embedding with varying scales of contexts, without addressing the global relevance of the image pair. Thus, the resultant alignment may suffer implausible voxel-wise mapping, where the prior affine transformation and landmark annotation are required to circumvent the trap of local minima.

Variants of transformers have gained great success in a group of tasks in natural language processing (NLP), including cross-language translation [25] and the question-answering [22]. Recently, the transformer has been extended to computer vision community, such as object detection [5], image recognition [10], and segmentation [7,23]. The transformer facilitates the global embedding of images by the relevance modeling of image words. Attention was utilized in various image processing tasks by highlighting salient feature regions and suppressing irrelevant ones [20]. Liao et al. [15] utilized an attention-driven hierarchical strategy and a greedy supervised approach in rigid CT registration. An auto-attention mechanism was introduced to multiple regions for reliable visual cues in the registration of X-ray and CT images [17]. Nevertheless, such attention schemes addressed the long-range dependencies of a single image or the rigid transformation, which can not effectively handle the cross-image semantic correspondence and deformable registration.

In response to these difficulties, we propose a dual transformer network (DTN) for diffeomorphic registration. The proposed approach exploits the self-attention scheme to model the inter- and intra-image global contextual relevances explicitly. The dual transformer conducts the relevance modeling and the feature enhancement on two kinds of image embedding for semantically meaningful correspondences of anatomical structures. The DTN consists of a learnable image embedding module, a cross-image relevance learning module, and a registration field inference module. The combinational embedding, taking the strength of both the low-level spatial features and the high-level contextual relevance-based enhancements, is used to predict the registration fields. One difficulty in unsupervised deformable registration is to identify the semantic correspondence between anatomical structures. The proposed DTN addresses the cross-image and global relevance to improve the discriminative power of image embedding for voxel-wise correspondence. We evaluate the proposed approach on the clinically obtained

brain MRI scans of the OASIS dataset [16] qualitatively and quantitatively. The atlas-based registration and segmentation demonstrate our model achieves performance improvements over the compared deep-learning-based methods. The main contributions of this work are as follows:

- We devise a novel dual transformer for volumetric diffeomorphic registration, facilitating the semantically meaningful correspondence of anatomical structures.
- We conduct the volume embedding enhancement for velocity field inference, taking advantage of both the CNN-based local features and the attention-based global and cross-image relevances.
- The proposed DTN has gained success in the diffeomorphic registration and atlas-based segmentation of multi-category anatomical structures.

## 2 Proposed Method

As shown in Fig. 1, the DTN can be stacked on an existing encoder-decoder network of the 3D U-net [8]. We present the dual transformer to take the strengths of both the CNN-based local features and the cross-image global contextual relevance-based enhancements for volumetric image embedding. Given the input moving and fixed volumetric images, $V_m, V_f \in \mathbb{R}^{h_0 \times w_0 \times d_0}$, the goal is to estimate the diffeomorphic registration field $\psi \in \mathbb{R}^{3h_0 \times w_0 \times d_0}$ for the one-to-one map and the atlas-based registration. The DTN has two branches to address the relevance learning on the volumetric embedding of separate one-channel images and the two-channel image concatenation. First, the DTN extracts the CNN-based low-level image embedding of both separate and the concatenated images. Second, The image embeddings are collapsed into sequences, which are fed to the dual

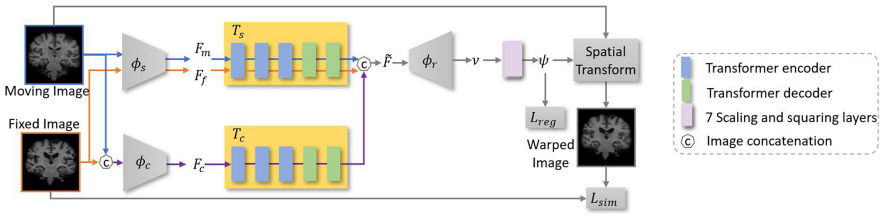


**Fig. 1.** Diffeomorphic registration by the proposed DTN. The framework is stacked on an existing CNN-based encoder-decoder network, such as 3D U-net, for volumetric image embedding and velocity field inference. The DTN consists of dual transformers, $T_s$ and $T_c$, to handle the cross-image global relevance learning on separate single-channel images and the image concatenation. The proposed DTN takes advantage of both the CNN-based low-level local features and the attention-based global relevances for diffeomorphic registration. (The skip connections of the 3D U-net are removed for the sake of the clarity.)

transformer for global relevance-based feature enhancement. Finally, the resultant features from two branches are concatenated together to infer the velocity field $\nu \in \mathbb{R}^{3h_0 \times w_0 \times d_0}$ and the registration field $\psi$.

### 2.1 Volumetric Image Embedding

The DTN utilizes the encoder of 3D U-net for embedding of separate and concatenated volumetric images. The first branch takes single-channel images as the input, and outputs features $F_m, F_f \in \mathbb{R}^{q_s \times h \times w \times d}$, and $F_{\{m,f\}} = \phi_s(V_{\{m,f\}}, \Theta_s)$. $q_s$ denotes the channel number and $\Theta_s$ network parameters. Aside from the embedding on single-channel volumetric image, we estimate the embedding of the image concatenation $[V_m\ V_f]$. The resultant $q_c$-channel feature $F_c \in \mathbb{R}^{q_c \times h \times w \times d}$, and $F_c = \phi_c([V_m\ V_f], \Theta_c)$. $\Theta_c$ denotes learnable network parameters.

### 2.2 Dual Transformer

We present a dual transformer to model the cross-volume dependencies to enhance the volumetric embedding, as shown in Fig. 1. We utilize an encoder-decoder transformer [6,24], consisting of concatenated three encoders and two decoders, to model the inter- and intra-volume relevance. The transformer encoder requires a sequence input, where the spatial dimensions of the features are collapsed into a vector with the resultant feature $\hat{F}_{\{m,f\}} \in \mathbb{R}^{q_s \times hwd}$ and $\hat{F}_c \in \mathbb{R}^{q_c \times hwd}$. A learnable position encoding is added to the feature sequence to retain the positional information as [24] (Fig. 2(a)).

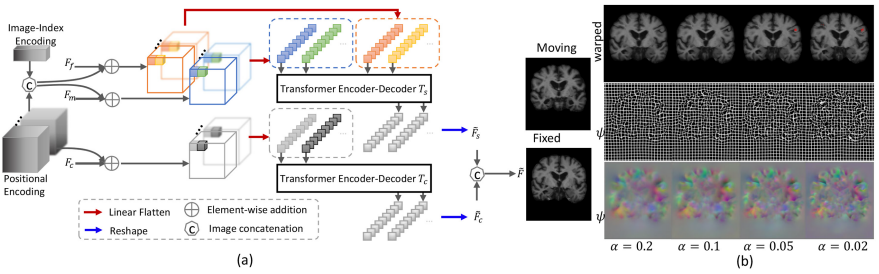


**Fig. 2.** (a) The dual transformer-based feature enhancements. (b) The warped moving images with varying $\alpha$. The registration fields $\psi$ are visualized in grid and color images. Red points on warped images indicate voxels with $|J(\psi)| \leq 0$. (Color figure online)

The first branch addresses the inter and intra-image relevances on the separate image embedding. Transformer $T_s$ handles the relevance of sequences from both the fixed and moving images, i.e., $\hat{F}_f$ and $\hat{F}_m$. $T_s$ takes $(\hat{F}_m, \hat{F}_f)$ and the positional encoding as the input and outputs $\tilde{F}_s \in \mathbb{R}^{q_s \times hwd}$. Here we use the image index encoding (Fig. 2 (a)) to indicate the fixed or moving image. Since

the queries and keys are defined as the embedding sequences from $\hat{F}_f$ and $\hat{F}_m$, $T_s$ models both the intra- and inter-image dependencies. The resultant feature embedding $\tilde{F}_s$ enhances the convolutional features with increasing receptive fields and the cross-image global relevance.

In the second branch, $T_c$ utilizes the self-attention scheme to model the global dependencies of the concatenated volumetric embedding $F_c$. The input of $T_c$ is $\hat{F}_c$ and the positional encoding. The resultant $\tilde{F}_c \in \mathbb{R}^{q_c \times hwd}$. Since the entangled embedding $F_c$ is computed from the image pair, $T_c$ also addresses the inter-volume feature enhancements. The final enhanced feature embedding $\tilde{F}$ is computed as a concatenation of features from both $T_s$ and $T_c$, and $\tilde{F} = \tilde{F}_s © \tilde{F}_c$. © denotes the image concatenation operator. We noticed the U-Net Transformer [19] had transformer modules in the skip connections for segmentation. The multiple transformers are promising to enhance features further, though enlarging the memory and computational complexity. We implemented the self-attention learning on the voxel sequence at the bottom of the U-net.

### 2.3 Diffeomorphic Registration

Since a diffeomorphism has the differentiable and invertible properties theoretically, the one-to-one mapping and the topology preservation are guaranteed in the diffeomorphic registration. We conduct the diffeomorphic registration and utilize the stationary velocity field $\nu_t, t \in [0, 1]$, which satisfies

$$d\psi_t/dt = \nu_t(\psi_t) = \nu_t \circ \psi_t. \tag{1}$$

The diffeomorphic deformation field $\psi^{(0)} = I$ is an identity transformation. The deformation field represented as a member of Lie algebra can be computed as the exponential of the velocity field [1]. The exponentiated velocity field $\psi^{(1)} = \exp(\nu)$ guarantees the mapping between images to be invertible. Given the enhanced feature $\tilde{F}$, the diffeomorphic registration decoder is used to infer the invertible deformation fields $\psi$. The decoder network parameterizes a nonlinear mapping function $\phi_r(\tilde{F}, \Theta_r) = \psi$, as a combination of a CNN-based decoder and scaling and squaring layers (Fig. 1). $\Theta_r$ denotes the parameters of the registration inference module.

### 2.4 Unsupervised Learning

The proposed DTN is optimized in an unsupervised manner by the metric space alignment. Given image pair $(V_m, V_f)$, the DTN estimates registration field $\psi$. The spatial transformer [13] is used to warp the moving image. The resultant warped moving image $V'_m = V_m \circ \psi$. The $l_1$-norm-based image similarity loss $L_{sim} = \|V'_m - V_f\|_1$. We apply the Frobenius norm-based smoothness regularizer on the velocity field, and $L_{reg} = \|\nabla\nu\|_F^2$. The loss function is defined as follows:

$$L = L_{sim} + \alpha L_{reg}, \tag{2}$$

where the hyperparameter $\alpha$ is used to balance the image similarity and the smoothness regularization on the velocity field. We optimize the parameters of the proposed DTN by minimizing the loss function.

## 3 Experiments

**Dataset and Metric.** We used the publicly available dataset of the OASIS with 425 T1-weighted brain MRI scans [16]. The MRI scans are re-sampled to a resolution of $256 \times 256 \times 256$ with an isotropic voxel size of $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$, and then cropped to $144 \times 160 \times 192$. We conducted the standard preprocessing for affine transformation and brain structure extraction using FreeSurfer [11] as [18]. The OASIS dataset provides brain segmentation with visual inspections, which are viewed as the ground truth in the evaluation process. The dataset was split into 256 and 150 scans for training and testing. The remaining 19 scans have been used for validation.

We evaluate the proposed approach using the Dice similarity coefficients (DSC), which measures the consistency between the ground truth segmentations and those estimated by the atlas-based registration. The negative Jacobian determinant $|J(\psi)| \leq 0$ is used to evaluate the registration fields. We compute the derivatives of the volumetric registration fields for the negative Jacobian determinant, which is related to structural folding without topology preservation and the violation of the diffeomorphic property. Generally, the low value of negative Jacobian determinant and the high value of the DSC suggest reliable diffeomorphic registration fields.

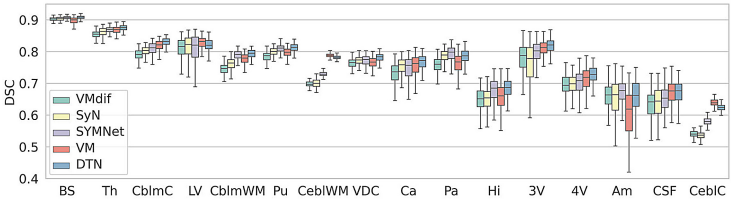


**Fig. 3.** Boxplots of DSCs on 16 anatomical structures, including brain stem (BS), thalamus (Th), cerebellum cortex (CblmC), lateral ventricle (LV), cerebellum white matter (CblmWM), putamen (Pu), cerebral white matter (CeblWM), ventral DC (VDC), caudate (Ca), pallidum (Pa), hippocampus (Hi), the 3rd ventricle (3V), the 4th ventricle (4V), amygdala (Am), CSF, and cerebral cortex (CeblC), by the SyN [2], the VM [4], the $\text{VM}_{dif}$ [9], the SYMNet [18], and the proposed DTN. The symmetric structures are combined into one in the plots.

**Implemental Details.** We built the DTN on the five-level 3D Unet, where the volumetric image embedding module and the registration field inference module are the encoder and the decoder of the U-net, respectively. The CNN-based feature embedding has $q_s = 128$ channels and $q_c = 96$ channels. The resultant embedding of the dual transformation has the resolution of $224 \times 9 \times 10 \times 12$. The hyperparameter $\alpha$ in the registration loss (2) is set to 0.1. The proposed framework is implemented using the PyTorch on a PC with a NVIDIA GTX TITAN xp GPU. The network parameters are optimized using the ADAM algorithm with a learning rate of $1e-4$. The momentums are set to 0.5 and 0.999.

The mini-batch consists of 1 image. The training takes 50 h with 400 epochs and 102.4 k iterations. The online testing takes 0.653 s.

### 3.1 Qualitative Assessment

The proposed DTN realizes the diffeomorphic registration of volumetric images. The dual transformer-based attention enhances the cross-image volumetric embedding, facilitating the diffeomorphic registration.

**Comparison with Sstate-of-the-Art.** We compare with the symmetric image normalization registration method (SyN) [2] and recent deep-learning-based registration models, including the VM [4], the diffeomorphic variant $\text{VM}_{dif}$ [9], and the SYMNet [18] (Figs. 3 and 4), using the publicly available implementations provided by authors with the suggested parameter setting. Both the VM [4] and the VMdif [9] use the MSE similarity loss, where the smooth regularization weight and the learning rate are set to 0.01 and 1e−4, respectively. The SYMNet [18] uses the NCC similarity loss, where the orientation consistency weight, the smooth regularization weight, the magnitude weight, and the learning rate are set to 1000, 10, 0.001, and 1e−4, respectively. We used the same data splitting for training, validation, and testing datasets in comparison. We use the ANTs toolbox [3] for the SyN implementations with the maximum iteration set to $[100, 100, 100]$ in the iterative optimization. The unsupervised VM predicts the displacement vector fields directly from the input image pair without the guarantee of the diffeomorphic property. The $\text{VM}_{dif}$, SYMnet, and the proposed DTN infer the velocity field for the diffeomorphic registration.

Table 1 reports the DSCs and the negative Jacobian determinant by the proposed DTN and the compared methods, including the SyN, the VM, the $\text{VM}_{dif}$, and the SYMNet. The proposed DTN is feasible to estimate the diffeomorphic registration field with low voxel numbers of the negative Jacobian determinants, facilitating performance improvements on the average DSC of anatomical structures. The deep learning-based methods, including the VM, the SYMNet, and the proposed DTN, achieve higher DSCs than the iterative optimization-based SyN. Similar to deep learning-based diffeomorphic registration models of the $\text{VM}_{dif}$ and the SYMNet, the proposed DTN is feasible to reduce the negative Jacobian determinant without sacrificing the atlas-based registration. Figure 3 illustrates the boxplot of DSCs on 16 anatomical structures. As we can see, the proposed DTN outperforms the compared diffeomorphic registration methods on 13 out of 16 structures. We further conducted the statistical significance tests. There existed statistically significant improvements with the p-values $< 0.05$ in terms of the average DSC in the t-test. The proposed DTN outperformed the compared transformer-free models with the p-values below 5e−15 (VM [4]), 5e−75 (VMdif [9]), and 5e−20 (SYMNet [18]). The dual transformer scheme effectively models the cross-image global relevances and enhances the volumetric image embedding for semantically meaningful correspondence. For instance, the boundaries of the lateral-ventricle are more consistent with the atlas than the compared methods

as shown in Fig. 4. Note that the proposed approach utilizes the basic diffeomorphic registration loss (2) on the image similarity and the regularization as the first-order gradients of the velocity field, without relying on the delicately designed regularizer.

**Table 1.** The DSC by the SyN [2], the VM [4], VM$_{dif}$ [9], SYMNet [18], and variants of our methods.

| | | SyN | VM | VM$_{dif}$ | SYMNet | DTN | DTN$_s$ | DTN$_c$ |
|---|---|---|---|---|---|---|---|---|
| DSC | Avg. | 0.740 | 0.756 | 0.733 | 0.755 | 0.769 | 0.763 | 0.764 |
| | Std. | 0.089 | 0.081 | 0.086 | 0.081 | 0.076 | 0.076 | 0.077 |
| $\lvert J(\psi)\rvert \leq 0$ | Avg. | 0 | 20604 | 2.84 | 1169 | 0.497 | 0.531 | 0.655 |
| | Std. | 0 | 3015 | 9.64 | 405 | 2.13 | 2.21 | 2.59 |

**Regularization.** We study the effect of the regularization on velocity fields. The deep learning-based diffeomorphic registration model is not guaranteed to produce invertible one-to-one map, due to computational errors in registration fields. Figure 2 (b) shows the voxels with negative Jacobian determinants with varying $\alpha$. The average voxel number with $|J(\psi)| \leq 0$ reduces from 1376 to 0 when $\alpha$ ranges from 0.02 to 0.2. By tuning the hyperparameter $\alpha$, the resultant registration field balances topology preservation with plausible registration accuracies.

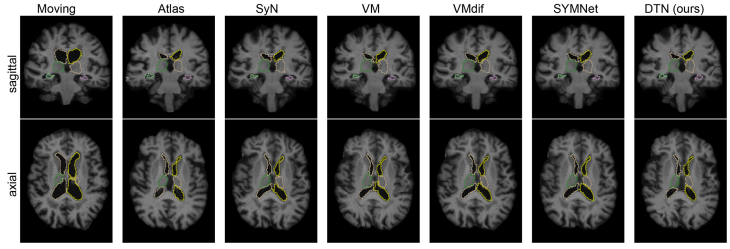


**Fig. 4.** The atlas-based registration of lateral-ventricle, thalamus, and hippocampus by the SyN [2], VM [4], VM$_{dif}$ [9], SYMNet [18], and ours.

**Ablation Study.** The proposed dual transformer network addressed the relevance modeling and the feature enhancement on two kinds of image embedding for semantically meaningful correspondences of anatomical structures. We conducted an ablation study to analyze the proposed dual transformer network (Table 1). The proposed DTN with the dual transformer architecture outperformed the DTN$_s$ and the DTN$_c$ with the only transformer of $T_s$ on the dependencies of the separate image embedding or $T_c$ on the relevance of the concatenated image embedding. We compared with the transformer-free deep diffeomorphic registration models, including the VMdif [9] and the SYMNet [18]. As

shown in Fig. 1, the encoder-decoder architecture with $\phi_c$ and $\phi_r$ followed the VMdif [9] when removing the dual transformer. As shown in Table 1, both the $\text{DTN}_s$ and $\text{DTN}_c$ are helpful to improve the registration accuracy compared with transformer-free models with the DSCs of $0.763 \pm 0.076$ and $0.764 \pm 0.077$, respectively. The transformer branches are complementary, where the proposed DTN with the dual transformers outperformed the $\text{DTN}_s$ and the $\text{DTN}_c$ with p-values below 1e−6 and 1e−4, respectively. We designed the dual architecture to identify the semantically meaningful correspondence of anatomical structures. The attention-based global and cross-image relevance enhanced volumetric embedding and improved the registration accuracy compared with transformer-free models (Table 1).

## 4 Conclusion

This paper presents the DTN for the volumetric diffeomorphic registration, taking advantage of both the CNN-based low-level features and the attention-based global and cross-image relevances for feature enhancements. The DTN explicitly models the long-range cross-image relevance in embedding learning to identify the semantically meaningful correspondence of anatomical structures. The qualitative and quantitative evaluations of the proposed approach are conducted on the OASIS dataset with clinically obtained brain MRI scans. The improvements on the atlas-based registration suggest that the dual transformer facilitates the semantically meaningful correspondence of anatomical structures.

**Acknowledgments.** This work was supported in part by National Natural Science Foundation of China under Grant 61876008 and 82071172, Beijing Natural Science Foundation under Grant 7192227, and Research Center of Engineering and Technology for Digital Dentistry, Ministry of Health.

## References

1. Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage **38**, 95–113 (2007)
2. Avants, B., Epstein, C., Grossman, M., Gee, J.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**(1), 26–41 (2008)
3. Avants, B., Tustison, N., Song, G., Cook, P., Klein, A., Gee, J.: A reproducible evaluation of ants similarity metric performance in brain image registration. NeuroImage **54**, 2033–2044 (2011)
4. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260 (2018)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. ArXiv abs/2005.12872 (2020)
6. Chen, H., et al.: Pre-trained image processing transformer. ArXiv abs/2012.00364 (2020)

7. Chen, J., et al.: TransuNet: transformers make strong encoders for medical image segmentation. ArXiv abs/2102.04306 (2021)
8. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424–432 (2016)
9. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 729–738. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_82
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. ArXiv abs/2010.11929 (2020)
11. Fischl, B.: Freesurfer. NeuroImage **62**, 774–781 (2012)
12. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vis. Appl. **31**, 1–18 (2020)
13. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NeurIPS, pp. 2017–2025 (2015)
14. Leow, A., et al.: Inverse consistent mapping in 3D deformable image registration: its construction and statistical properties. Inf. Process. Med. Imaging **19**, 493–503 (2005)
15. Liao, R., et al.: An artificial agent for robust image registration. In: AAAI (2017)
16. Marcus, D., Wang, T.H., Parker, J., Csernansky, J., Morris, J., Buckner, R.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**, 1498–1507 (2007)
17. Miao, S., et al.: Dilated FCN for multi-agent 2D/3D medical image registration. ArXiv abs/1712.01651 (2018)
18. Mok, T.C.W., Chung, A.C.S.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4643–4652 (2020)
19. Petit, O., Thome, N., Rambour, C., Soler, L.: U-net transformer: self and cross attention for medical image segmentation. ArXiv abs/2103.06104 (2021)
20. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
21. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. IEEE Trans. Med. Imaging **32**(7), 1153–1190 (2013)
22. Tan, H.H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. In: EMNLP/IJCNLP (2019)
23. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.: Medical transformer: gated axial-attention for medical image segmentation (2021)
24. Vaswani, A., et al.: Attention is all you need. ArXiv abs/1706.03762 (2017)
25. Wang, Q., et al.: Learning deep transformer models for machine translation. ArXiv abs/1906.01787 (2019)
26. Zhang, J.: Inverse-consistent deep networks for unsupervised deformable image registration. ArXiv abs/1809.03443 (2018)
27. Zhang, Y., Pei, Y., Guo, Y., Ma, G., Xu, T., Zha, H.: Fully convolutional network for consistent voxel-wise correspondence. In: AAAI (2020)