ModeT: Learning Deformable Image Registration via Motion Decomposition Transformer

Haiqiao Wang, Dong Ni, and Yi Wang(✉)

Smart Medical Imaging, Learning and Engineering (SMILE) Lab, Medical UltraSound Image Computing (MUSIC) Lab, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China
onewang@szu.edu.cn

Abstract. The Transformer structures have been widely used in computer vision and have recently made an impact in the area of medical image registration. However, the use of Transformer in most registration networks is straightforward. These networks often merely use the attention mechanism to boost the feature learning as the segmentation networks do, but do not sufficiently design to be adapted for the registration task. In this paper, we propose a novel motion decomposition Transformer (ModeT) to explicitly model multiple motion modalities by fully exploiting the intrinsic capability of the Transformer structure for deformation estimation. The proposed ModeT naturally transforms the multi-head neighborhood attention relationship into the multi-coordinate relationship to model multiple motion modes. Then the competitive weighting module (CWM) fuses multiple deformation sub-fields to generate the resulting deformation field. Extensive experiments on two public brain magnetic resonance imaging (MRI) datasets show that our method outperforms current state-of-the-art registration networks and Transformers, demonstrating the potential of our ModeT for the challenging non-rigid deformation estimation problem. *The benchmarks and our code are publicly available at* https://github.com/ZAX130/SmileCode.

Keywords: Deformable image registration · Motion decomposition · Transformer · Attention · Pyramid structure

1 Introduction

Deformable image registration has always been an important focus in the society of medical imaging, which is essential for the preoperative planning, intraoperative information fusion, disease diagnosis and follow-ups [10,23]. The deformable registration is to solve the non-rigid deformation field to warp the moving image, so that the warped image can be anatomically similar to the fixed image. Let $I_f, I_m \in \mathbb{R}^{H\times W\times L}$ be the fixed and moving images (H, W, L denote image size), in the deep-learning-based registration paradigm, it is often necessary to employ a spatial transformer network (STN) [13] to apply the estimated sampling grid $G \in \mathbb{R}^{H\times W\times L\times 3}$ to the moving image, where G is obtained by adding the regular grid and the deformation field. For any position $p \in \mathbb{R}^3$ in the sampling

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 740–749, 2023.
https://doi.org/10.1007/978-3-031-43999-5_70

grid, $G(p)$ represents the corresponding relation, which means that the voxel at position p in the fixed image corresponds to the voxel at position $G(p)$ in the moving image. That is to say, image registration can be understood as finding the corresponding voxels between the moving and fixed images, and converting this into the relative positional relationship between voxels, which is very similar to the calculation method of Transformer [8].

Transformers have been successfully used in the society of computer vision and have recently made an impact in the field of medical image computing [11,17]. In medical image registration, there are also several related studies that employ Transformers to enhance network structures to obtain better registration performance, such as Transmorph [5], Swin-VoxelMorph [26], ViT-V-Net [6], etc. The use of Transformer in these networks, however, often merely leverages the self-attention mechanism in Transformers to boost the feature learning (the same as the segmentation tasks do), but does not sufficiently design for the registration tasks. Some other methods use cross-attention to model the corresponding relationship between moving and fixed images, such as Attention-Reg [22] and Xmorpher [21]. The cross-attention Transformer (CAT) module is used in the bottom layer of Attention-Reg [22] and each layer in Xmorpher [21] to establish the relationship between the features of moving and fixed images. However, the usage of Transformer in [21,22] is still limited to improving the feature learning, with no additional consideration given to the relationship between the attention mechanism and the deformation estimation. Furthermore, due to the large network structure of [21], only small windows can be created for similarity calculation, which may result in performance degradation. Few studies consider the relationship between attention and deformation estimation, such as Coordinate Translator [18] and Deformer [4]. Deformer [4] uses the calculation mode of multiplication of attention map and Value matrix in Transformer to weight the predicted basis to generate the deformation field, but its attention map calculation is only the concatenation and projection of moving and fixed feature maps, without using similarity calculation part. Coordinate Translator [18] calculates the matching score of the fixed feature map and the moving feature map. Then the computed scores are employed to re-weight the deformation field. However, for feature maps with coarse-level resolution, a voxel often has multiple possibilities of different motion modes [25], which is not considered in [18]. Traditional methods have explored multiple modes of deformations, e.g., probabilistic registration [12], to improve the performance.

In this study, we propose a novel motion decomposition Transformer (ModeT) to explicitly model multiple motion modalities by fully exploiting the intrinsic capability of the Transformer structure for deformation estimation. Experiments on two public brain magnetic resonance imaging (MRI) datasets demonstrate our method outcompetes several cutting-edge registration networks and Transformers. The main contributions of our work are summarized as follows:

- We propose to leverage the Transformer structure to naturally model the correspondence between images and convert it into the deformation field,

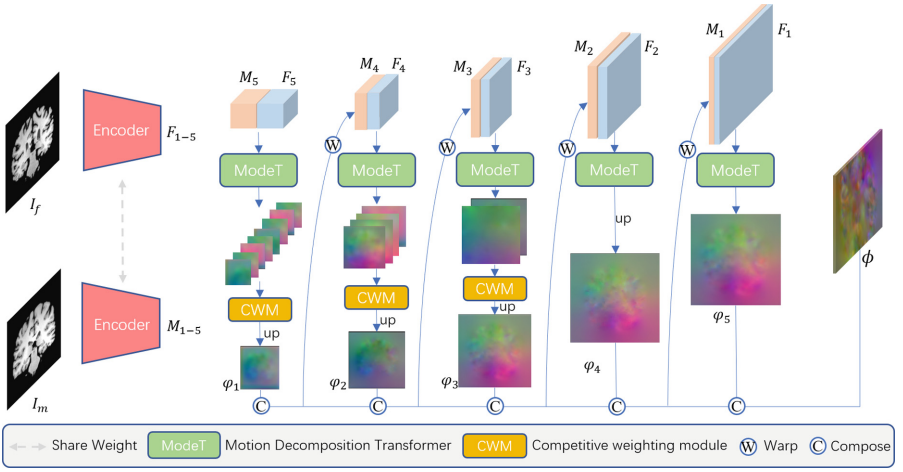

Fig. 1. Illustration of the proposed deformable registration network. The encoder takes the fixed image I_f and moving image I_m as input to extract hierarchical features F_1-F_5 and M_1-M_5. The motion decomposition Transformer (ModeT) is used to generate multiple deformation sub-fields and the competitive weighting module (CWM) fuses them. Finally the decoding pyramid outputs the total deformation field ϕ.

thus explicitly separating the two tasks of feature extraction and deformation estimation in deep-learning-based registration networks in which to make the registration procedure more sensible.

- The proposed ModeT makes full use of the multi-head neighborhood attention mechanism to efficiently model multiple motion modalities, and then the competitive weighting module (CWM) fuses multiple deformation sub-fields in a competitive way, which can improve the interpretability and consistency of the resulting deformation field.
- The pyramid structure is employed for feature extraction and deformation propagation, and is beneficial to reduce the scope of attention calculation required for each level.

2 Method

2.1 Network Overview

The proposed deformable registration network is illustrated in Fig. 1. We employ a pyramidal registration structure, which has the advantage of reducing the scope of attention calculation required at each decoding level and therefore alleviating the computational consumption. Given the fixed image I_f and moving image I_m as input, the encoder extracts hierarchical features using a 5-layer convolutional block, which doubles the number of channels in each layer. This generates two sets of feature maps F_1, F_2, F_3, F_4, F_5 and M_1, M_2, M_3, M_4, M_5. The feature

maps M_5 and F_5 are sent into the ModeT to generate multiple deformation sub-fields, and then the generated deformation sub-fields are input into the CWM to obtain the fused deformation field φ_1 of the coarsest decoding layer as the initial of the total deformation field ϕ. The moving feature map M_4 is deformed using ϕ, and the deformed moving feature map is fed into the ModeT along with F_4 to generate multiple sub-fields, which are input into the CWM to get φ_2. Then φ_2 is compounded with previous total deformation field to generate the updated ϕ. The feature maps M_3 and F_3 go through the similar operations. As the decoding feature maps become finer, the number of motion modes at position p decreases, along with the number of attention heads we need to model. At the F_2/M_2 and F_1/M_1 levels, we no longer generate multiple deformation sub-fields, i.e., the number of attention heads in ModeT is 1. Finally, the obtained total deformation field ϕ is used to warp I_m to obtain the registered image.

To guide the network training, the normalized cross correlation $\mathcal{L}_{\text{ncc}}$ [19] and the deformation regularization $\mathcal{L}_{\text{reg}}$ [3] is used:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{ncc}}(I_f, I_m \circ \phi) + \lambda \mathcal{L}_{\text{reg}}(\phi), \tag{1}$$

where $\circ$ is the warping operation, and λ is the weight of the regularization term.

2.2 Motion Decomposition Transformer (ModeT)

In deep-learning-based registration networks, a position p in the low-resolution feature map contains semantic information of a large area in the original image and therefore may often have multiple possibilities of different motion modalities. To model these possibilities, we employ a multi-head neighborhood attention mechanism to decompose different motion modalities at low-resolution level. The illustration of the motion decomposition is shown in Fig. 2.

Let $F, M \in \mathbb{R}^{c \times h \times w \times l}$ stand for the fixed and moving feature maps from a specific level of the hierarchical encoder, where h, w, l denote feature map size and c is the channel number. The feature maps F and M go through linear projection ($proj$) and LayerNorm (LN) [2] to get Q ($query$) and K (key):

$$\begin{aligned} Q = LN(proj(F)), \quad K = LN(proj(M)), \\ Q = \{Q^{(1)}, Q^{(2)}, \ldots, Q^{(S)}\}, \\ K = \{K^{(1)}, K^{(2)}, \ldots, K^{(S)}\}, \end{aligned} \tag{2}$$

where the projection operation is shared weight, and the weight initialization is sampled from $N(0, 1e^{-5})$, the bias is initialized to 0. The Q and K are then divided according to channels, and S represents the number of divided heads.

We then calculate the neighborhood attention map. We use $c(p)$ to denote the neighborhood of voxel p. For a neighborhood of size $n \times n \times n$, $||c(p)|| = n^3$. The neighborhood attention map of multiple heads is obtained by:

$$NA(p, s) = softmax(Q_p^{(s)} \cdot K_{c(p)}^{(s)T} + B^{(s)}), \tag{3}$$

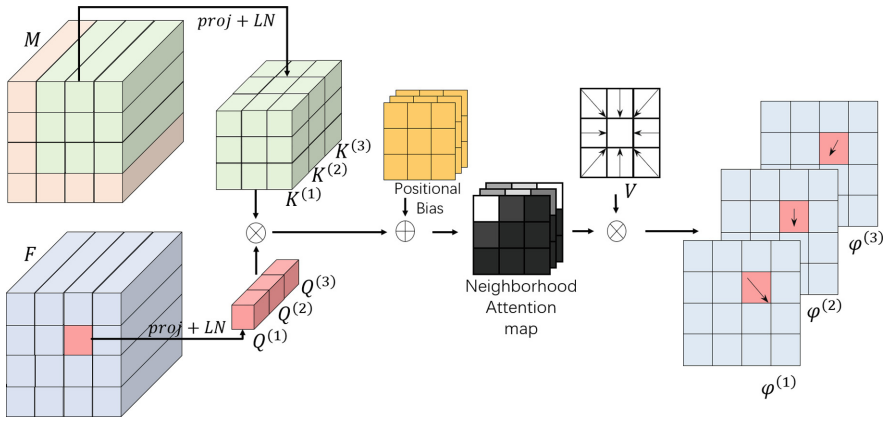

Fig. 2. Illustration of the proposed motion decomposition Transformer, which employs the multi-head neighborhood attention mechanism to decompose different motion modalities. ($S = 3$ in this illustration)

where $B \in \mathbb{R}^{S\times n\times n\times n}$ is a learnable relative positional bias, initialized to all zeros. We pad the moving feature map with zeros to calculate boundary voxels because the registration task sometimes requires voxels outside the field-of-view to be warped. Equation (3) shows how the neighborhood attention is computed for the s-th head at position p, so that the semantic information of voxels on low resolution can be decomposed to compute similarity one by one, in preparation for modeling different motion modalities. Moreover, the neighborhood attention operation narrows the scope of attention calculation to reduce the computational effort, which is very friendly to volumetric processing.

The next step is to obtain the multiple sub-fields at this level by computing the regular displacement field weighted via the neighborhood attention map:

$$\varphi_p^{(s)} = NA(p, s)V, \tag{4}$$

where $\varphi^{(s)} \in \mathbb{R}^{h\times w\times l\times 3}$, $V \in \mathbb{R}^{n\times n\times n}$, and V $(value)$ represents the relative position coordinates for the neighborhood centroid, which is not learned so that the multi-head attention relationship can be naturally transformed into a multi-coordinate relationship. With the above steps, we obtain a series of deformation sub-fields for this level:

$$\varphi^{(1)}, \varphi^{(2)}, \ldots, \varphi^{(S)} \tag{5}$$

2.3 Competitive Weighting Module (CWM)

Multiple low-resolution deformation fields need to be reasonably fused when deforming a high-resolution feature map. As shown in Fig. 3, we first upsample these deformation sub-fields, then convolve them in three layers to get the score of each sub-field, and use softmax to compete the motion modality for each

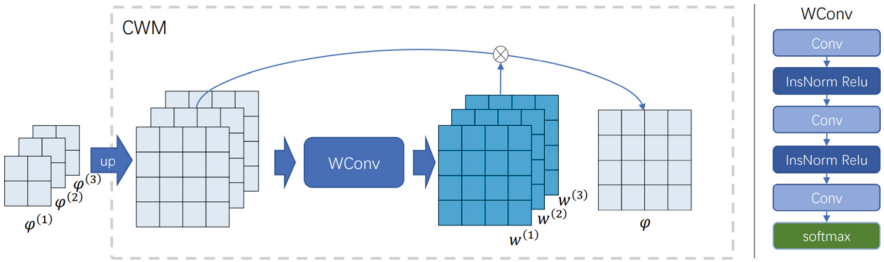

Fig. 3. Illustration of the proposed competitive weighting module (CWM).

voxel. The convolution uses $3 \times 3 \times 3$ convolution rather than direct projection because deformation fields often require correlation of adjacent displacements to determine if they are reasonable. We formulate above competitive weighting operation to obtain the deformation field φ at this level as follows:

$$\begin{aligned} w^{(1)}, w^{(2)}, \ldots, w^{(S)} &= WConv(cat(\varphi^{(1)}, \varphi^{(2)}, \ldots, \varphi^{(S)})), \\ \varphi &= w^{(1)}\varphi^{(1)} + w^{(2)}\varphi^{(2)} +, \ldots, + w^{(S)}\varphi^{(S)}, \end{aligned} \tag{6}$$

where $w^{(s)} \in \mathbb{R}^{h \times w \times l}$, and $\varphi^{(s)}$ has already been upsampled. $WConv$ represents the ConvBlock used to calculate weights, as shown in the right part of Fig. 3.

3 Experiments

Datasets. Experiments were carried on two public brain MRI datasets, including LPBA [20] and Mindboggle [16]. For LPBA, each MRI volume contains 54 manually labeled region-of-interests (ROIs). All volumes in LPBA were rigidly pre-aligned to mni305. 30 volumes (30×29 pairs) were employed for training and 10 volumes (10×9 pairs) were used for testing. For Mindboggle, each volume contains 62 manually labeled ROIs. All volumes in Mindboggle were affinely aligned to mni152. 42 volumes (42×41 pairs from the NKI-RS-22 and NKI-TRT-20 subsets) were employed for training, and 20 volumes from OASIS-TRT-20 (20×19 pairs) were used for testing. All volumes were pre-processed by min-max normalization, and skull-stripping using FreeSurfer [9]. The final size of each volume was $160 \times 192 \times 160$ after a center-cropping operation.

Evaluation Metrics. To quantitatively evaluate the registration performance, Dice score (DSC) [7] was calculated as the primary similarity metric to evaluate the degree of overlap between corresponding regions. In addition, the average symmetric surface distance (ASSD) [24] was evaluated, which can reflect the similarity of the region contours. The quality of the predicted deformation ϕ was assessed by the percentage of voxels with non-positive Jacobian determinant (i.e., folded voxels). All above metrics were calculated in 3D. A better registration shall have larger DSC, and smaller ASSD and Jacobian.

Table 1. The numerical results of different registration methods on two datasets.

	Mindboggle (62 ROIs)			LPBA (54 ROIs)		
	DSC (%)	ASSD	$\%\lvert J_\phi\rvert \leq 0$	DSC (%)	ASSD	$\%\lvert J_\phi\rvert \leq 0$
`SyN` [1]	56.7 ± 1.5	1.38 ± 0.09	$< 0.00001\%$	70.1 ± 6.2	1.72 ± 0.12	$< 0.0004\%$
`VM` [3]	56.0 ± 1.6	1.49 ± 0.11	$< 1\%$	64.3 ± 3.2	2.03 ± 0.21	$< 0.7\%$
`TM` [5]	60.7 ± 1.5	1.35 ± 0.10	$< 0.9\%$	67.0 ± 3.0	1.90 ± 0.20	$< 0.6\%$
`I2G` [18]	59.8 ± 1.3	1.30 ± 0.07	$< 0.03\%$	71.0 ± 1.4	1.64 ± 0.10	$< 0.01\%$
`PR++` [14]	61.1 ± 1.4	1.34 ± 0.10	$< 0.5\%$	69.5 ± 2.2	1.76 ± 0.17	$< 0.2\%$
`XM` [21]	53.6 ± 1.5	1.46 ± 0.09	$< 1\%$	66.3 ± 2.0	1.92 ± 0.15	$< 0.1\%$
`DMR` [4]	60.6 ± 1.4	1.34 ± 0.09	$< 0.7\%$	69.2 ± 2.4	1.79 ± 0.18	$< 0.4\%$
`Ours`	$\mathbf{62.8 \pm 1.2}$	$\mathbf{1.22 \pm 0.07}$	$< 0.03\%$	$\mathbf{72.1 \pm 1.4}$	$\mathbf{1.58 \pm 0.11}$	$< 0.007\%$

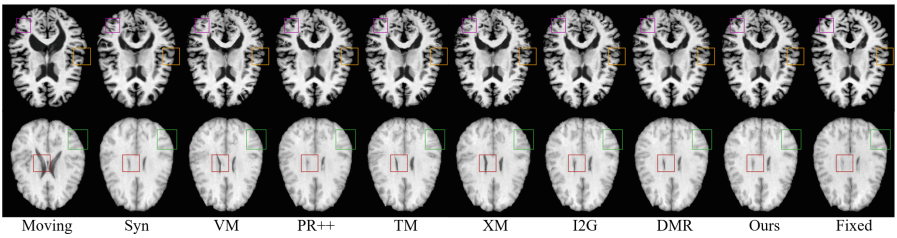

Fig. 4. Visualized registration results from different methods on Mindboggle (top row) and LPBA (bottom row).

Implementation Details. Our method was implemented with PyTorch, using a GPU of NVIDIA Tesla V100 with 32GB memory. The regularization term λ and neighborhood size n were set as 1 and 3. For the encoder part, we used the same convolution structure as [18]. In the pyramid decoder, from coarse to fine, the number of attention heads were set as $8, 4, 2, 1, 1$, respectively. We used 6 channels for each attention head. The Adam optimizer [15] with a learning rate decay strategy was employed as follows:

$$lr_m = lr_{init} \cdot (1 - \frac{m-1}{M})^{0.9}, m = 1, 2, ..., M \tag{7}$$

where lr_m represents the learning rate of m-th epoch and $lr_{init} = 1e-4$ represents the learning rate of initial epoch. We set the batch size as 1, M as 30 for training. In the inference phase, our method averagely took 0.56 second and 9GB memory to register a volume pair of size $160 \times 192 \times 160$.

Comparison Methods. We compared our method with several state-of-the-art registration methods: (1) `SyN` [1]: a classical traditional approach, using the *SyNOnly* setting in ANTS. (2) `VoxelMorph(VM)` [3]: a popular single-stage registration network. (3) `TransMorph(TM)` [5]: a single-stage registration network

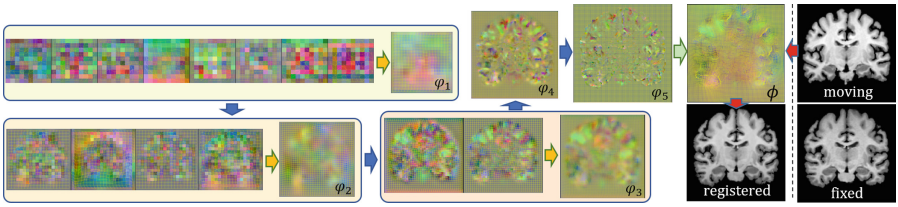

Fig. 5. Visualization of the generated multi-level deformation fields (φ_1-φ_5) to register one image pair. At low-resolution levels, multiple deformation sub-fields are decomposed to effectively model different motion modalities.

with SwinTransformer enhanced encoder. (4) `PR++` [14]: a pyramid registration network using 3D correlation layer. (5) `XMorpher(XM)` [21]: a registration network using CAT modules for each level of encoder and decoder. (6) `Im2grid(I2G)` [18]: a pyramid network using a coordinate translator. (7) `DMR` [4]: a registration network using a Deformer and a multi-resolution refinement module.

Quantitative and Qualitative Analysis. The numerical results of different methods on datasets Mindboggle and LPBA are reported in Table 1. It can be observed that our method consistently attained the best registration accuracy with respect to DSC and ASSD metrics. For the DSC results, our method surpassed the second-best networks by 1.7% and 1.1% on Mindboggle and LPBA, respectively. We further investigated the statistical significance of our method over comparison methods on DSC and ASSD metrics, by conducting the paired and two-sided Wilcoxon signed-rank test. The null hypotheses for all pairs (our method *v.s.* other method) were not accepted at the 0.05 level. As a result, our method can be regarded as significantly better than all comparison methods on DSC and ASSD metrics. Table 1 also lists the percentage of voxels with non-positive Jacobian determinant ($\%|J_\phi| \leq 0$). Our method achieved satisfactory performance, which was the best among all deep-learning-based networks.

Figure 4 visualizes the registered images from different methods on two datasets. Our method generated more accurate registered images, and internal structures can be consistently preserved using our method. Figure 5 takes the registration of one image pair as an example to show the multi-level deformation fields generated by our method. Our ModeT effectively modeled multiple motion modalities and our CWM fused them together at low-resolution levels. The final deformation field ϕ accurately warped the moving image to registered with the fixed image.

4 Conclusion

We present a motion decomposition Transformer (ModeT) to naturally model the correspondence between images and convert this into the deformation field,

which improves the interpretability of the deep-learning-based registration network. The proposed ModeT employs the multi-head neighborhood attention mechanism to identify various motion patterns of a voxel in the low-resolution feature map. Then with the help of competitive weighting module and pyramid structure, the motion modes contained in a voxel can be gradually fused and determined in the coarse-to-fine pyramid decoder. The experimental results have proven the superior performance of the proposed method. In our future study, we attempt to implement our ModeT in a more efficient way, and also investigate more effective fusion strategy to combine the displacement field from multiple attention heads.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grants 62071305, 61701312, 81971631 and 62171290, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011241, and in part by the Shenzhen Science and Technology Program (No. SGDX 20201103095613036).

References

1. 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)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
4. Chen, J., et al.: Deformer: towards displacement field learning for unsupervised medical image registration. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 141–151. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_14
5. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: transformer for unsupervised medical image registration. Med. Image Anal. **82**, 102615 (2022)
6. Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y.: ViT-V-Net: vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468 (2021)
7. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
9. Fischl, B.: FreeSurfer. NeuroImage **62**(2), 774–781 (2012)
10. Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: Deep learning in medical image registration: a review. Phys. Med. Biol. **65**(20), 20TR01 (2020)
11. He, K., et al.: Transformers in medical image analysis. Intell. Med. **3**(1), 59–78 (2023)
12. Heinrich, M.P., Simpson, I.J., Papież, B.W., Brady, S.M., Schnabel, J.A.: Deformable image registration by combining uncertainty estimates from supervoxel belief propagation. Med. Image Anal. **27**, 57–71 (2016)

13. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
14. Kang, M., Hu, X., Huang, W., Scott, M.R., Reyes, M.: Dual-stream pyramid registration network. Med. Image Anal. **78**, 102379 (2022)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Klein, A., Tourville, J.: 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. **6**, 171 (2012)
17. Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K.: Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. Med. Image Anal. **85**, 102762 (2023)
18. Liu, Y., Zuo, L., Han, S., Xue, Y., Prince, J.L., Carass, A.: Coordinate translator for learning deformable medical image registration. In: Li, X., Lv, J., Huo, Y., Dong, B., Leahy, R.M., Li, Q. (eds.) MMMI 2022. LNCS, vol. 13594, pp. 98–109. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18814-5_10
19. Rao, Y.R., Prathapani, N., Nagabhooshanam, E.: Application of normalized cross correlation to image registration. Int. J. Res. Eng. Technol. **3**(5), 12–16 (2014)
20. Shattuck, D.W., et al.: Construction of a 3D probabilistic atlas of human cortical structures. Neuroimage **39**(3), 1064–1080 (2008)
21. Shi, J., et al.: Xmorpher: full transformer for deformable medical image registration via cross attention. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 217–226. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_21
22. Song, X., et al.: Cross-modal attention for MRI and ultrasound volume registration. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 66–75. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_7
23. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. IEEE Trans. Med. Imaging **32**(7), 1153–1190 (2013)
24. Taha, A.A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging **15**(1), 1–28 (2015)
25. Zheng, J.Q., Wang, Z., Huang, B., Lim, N.H., Papiez, B.W.: Residual aligner network. arXiv preprint arXiv:2203.04290 (2022)
26. Zhu, Y., Lu, S.: Swin-VoxelMorph: a symmetric unsupervised learning model for deformable medical image registration using swin transformer. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 78–87. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_8