Swin-VoxelMorph: A Symmetric Unsupervised Learning Model for Deformable Medical Image Registration Using Swin Transformer

Yongpei Zhu(✉) and Shi Lu

Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
zhuyp20@mails.tsinghua.edu.cn

Abstract. Deformable medical image registration is widely used in medical image processing with the invertible and one-to-one mapping between images. While state-of-the-art image registration methods are based on convolutional neural networks, few attempts have been made with Transformers which show impressive performance on computer vision tasks. Existing 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. These methods also ignore the topology preservation and invertibility of the transformation although they achieve fast image registration. In this paper, we propose a novel, symmetric unsupervised learning network Swin-VoxelMorph based on the Swin Transformer which minimizes the dissimilarity between images and estimates both forward and inverse transformations simultaneously. Specifically, we propose 3D Swin-UNet, which applies hierarchical Swin Transformer with shifted windows as the encoder to extract context features. And a symmetric Swin Transformer-based decoder with patch expanding layer is designed to perform the up-sampling operation to estimate the registration fields. Besides, our objective loss functions can guarantee substantial diffeomorphic properties of the predicted transformations. We verify our method on two datasets including ADNI and PPMI, and it achieves state-of-the-art registration accuracy while maintaining desirable diffeomorphic properties.

Keywords: Medical image registration · Swin transformer · Swin-VoxelMorph · Diffeomorphic registration fields

1 Introduction

Deformable image registration is crucial in a variety of medical imaging analysis, which is aimed to estimate the appropriate non-linear transformation to align

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16446-0_8.

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13436, pp. 78–87, 2022.
https://doi.org/10.1007/978-3-031-16446-0_8

a pair of images. The deformable registration produces the nonlinear voxel-wise mapping between images, which facilitates the atlas-based annotation, statistical shape analysis, and shape comparison of anatomical structures. To accomplish the task of deformable registration effectively, we need to infer the semantic correspondence of fine-grained structures. The volumetric images vary in shapes and scales, so it is a challenging problem to identify the real matching anatomical structures.

It is known that traditional deformable registration has high computational cost due to iterative optimization of large-scale parameters [13]. Several deep learning-based approaches are proposed to train convolutional neural networks (CNNs) that map input pairs to output deformation [6,15,16,19,21–23,25,27]. 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 [5,6], for the invertible and the diffeomorphic transformations. Moreover, the symmetric registration infers a pair of diffeomorphic maps regarding the middle of the geodesic path [6]. However, on the one hand, although these methods achieve fast registration and comparable registration performance, the substantial diffeomorphic properties of the registration field are not guaranteed, including topology-preservation and the invertibility of the transformation. On the other hand, these methods conduct an inference directly from the CNN-based low-level local embedding without considering 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. What's more, registration is the process of establishing such correspondence by comparing different parts of the moving to the fixed image. Unlike CNNs, one point is that the self-attention mechanisms in a Transformer have an unlimited size effective receptive field. A CNN has a narrow field of view: it performs convolution locally, and its field of view grows in proportion to the CNN's depth, the shallow layers have a relatively small receptive field, limiting the CNN's ability to associate the distant parts between two images [28]. The U-Net (or other multi-scale pyramid modules) was proposed to overcome this limitation by introducing down- and up-sampling operations. However, several problems remain: (1) The receptive fields of the first several layers are still restricted by the convolution kernel size, and the global information of an image can only be viewed at the deeper layers of the network. (2) As the convolutional layers deepen, the impact from far-away voxels decays quickly [29]. However, Transformer is capable of handling such issues and focusing on the parts that need deformation.

Recently, the transformer has been extended to computer vision tasks, such as object detection [14], image recognition [12], and segmentation [10,11]. 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 [9]. Liao et al. [8] adopted an attention-driven hierarchical strategy and a greedy supervised method 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 [7]. However, such attention schemes solved the long-range dependencies of a single image or the rigid transformation, which can not effectively deal with the cross-image semantic correspondence and deformable registration.

To solve these difficulties, we propose a symmetric unsupervised learning network Swin-VoxelMorph which can minimize the dissimilarity between images and estimates both forward and inverse transformations simultaneously. On the one hand, the proposed method exploits the self-attention scheme to model the inter- and intra-image global contextual relevances explicitly. The transformer model conducts the relevance modeling and the feature enhancement on two kinds of image embedding for semantically meaningful correspondences of anatomical structures. The learnable embedding module is used to predict the registration fields, which takes the strength of both the low-level spatial features and the high-level contextual relevance-based enhancements. One difficulty in unsupervised deformable registration is to identify the semantic correspondence between anatomical structures. The proposed model addresses the cross-image and global relevance to improve the discriminative ability of image embedding for voxel-wise correspondence. To the best of our knowledge, we are the first to explicitly exploit Swin Transformer for deformable medical image registration. On the other hand, driven by [6], we also use several objective loss functions which can guarantee substantial diffeomorphic properties of the predicted transformations. We verify our method on two datasets including ADNI and PPMI, and obtain excellent improvement on magnetic resonance image (MRI) registration with higher average Dice scores and better diffeomorphic registration fields (lower non-positive Jacobian locations) compared with state-of-the-art methods. The contributions of this paper are two folds:

(1) We are the first to propose 3D Swin-UNet, a pure Transformer-based 3D U-shaped Encoder-Decoder network, to explicitly exploit Swin Transformer for deformable medical image registration. The learnable embedding module, 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.
(2) Our objective functions including orientation and inverse consistency constraint can guarantee the topology-preservation and inverse consistency of the predicted transformations.

2 Method

2.1 Network Structures

Deformable image registration refers to the process of warping one (moving) image to align with another (fixed) image to maximize the similarity between the registered images. Figure 1 shows the overview of our proposed symmetric architecture Swin-VoxelMorph. Let M, $F \in \mathbb{R}^{H \times W \times D}$ be moving and fixed image volumes. The optimization problem aims to minimize the dissimilarity of the fixed image F and warped image $M(\boldsymbol{\phi})$ while maintaining a smooth displacement field $\boldsymbol{\phi}$. We take M, F as inputs, and compute $\boldsymbol{\phi}$ based on parameter θ (the kernels of the convolutional layers) using the proposed transformer-based architecture 3D Swin-UNet which is shown in Fig. 2, where $\boldsymbol{\phi} = \boldsymbol{Id}+\boldsymbol{u}$, $\boldsymbol{u}$ denotes a flow field of displacement vectors, and $\boldsymbol{Id}$ denotes the identity. We warp M to $M(\boldsymbol{\phi}_{MF})$ and F to $F(\boldsymbol{\phi}_{FM})$ using differentiable spatial transformation functions.

The deep neural network (g_θ) here is built from the Swin-UNet [10] and consists of a 3-level hierarchical encoder-decoder with skip connections, which concatenates M and F into a 2-channel 3D image as the input. The basic unit of 3D Swin-UNet is Swin Transformer block [4]. In the encoder, the medical images are split into non-overlapping patches with patch size of $4 \times 4 \times 4$ to transform the inputs into sequence embeddings. By such partition approach, the feature dimension of each patch becomes to $4 \times 4 \times 4 \times 2 = 128$. Furthermore, a linear embedding layer is applied to projected feature dimension into arbitrary dimension (represented as C). The transformed patch tokens will generate the hierarchical feature representations by passing through several Swin Transformer blocks and patch merging layers. Specifically, Swin Transformer block is responsible for feature representation learning and patch merging layer is responsible for down-sampling and increasing dimension. Inspired by 3D U-Net [3], we design a symmetric transformer-based decoder which is composed of Swin Transformer block and patch expanding layer. The extracted context features are fused with multi-scale features from encoder via skip connections to complement the loss of spatial information caused by down-sampling. In contrast to patch merging layer, a patch expanding layer is specially designed to perform up-sampling. The patch expanding layer reshapes feature maps of adjacent dimensions into a large feature maps with 2× up-sampling of resolution. Finally, the last patch expanding layer is used to perform 4× up-sampling to restore the resolution of the feature maps to the input resolution ($D \times W \times H$), and then a linear projection layer is applied on these up-sampled features to estimate two registration fields $\boldsymbol{\phi}_{MF}$ and $\boldsymbol{\phi}_{FM}$.

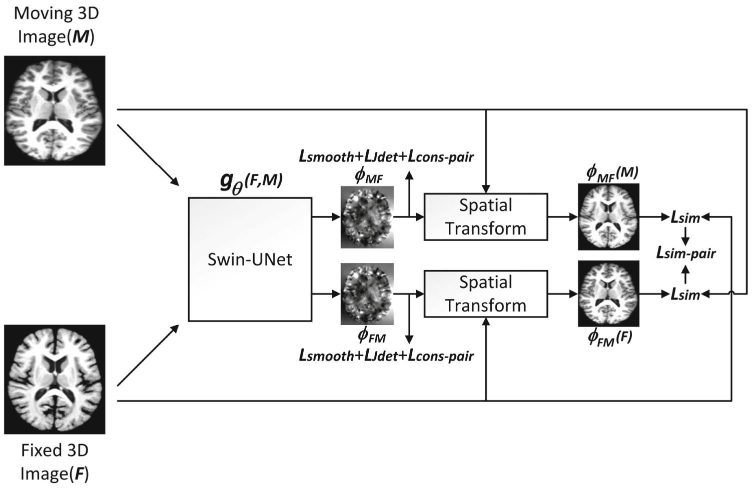

Fig. 1. Overview of our method Swin-VoxelMorph.

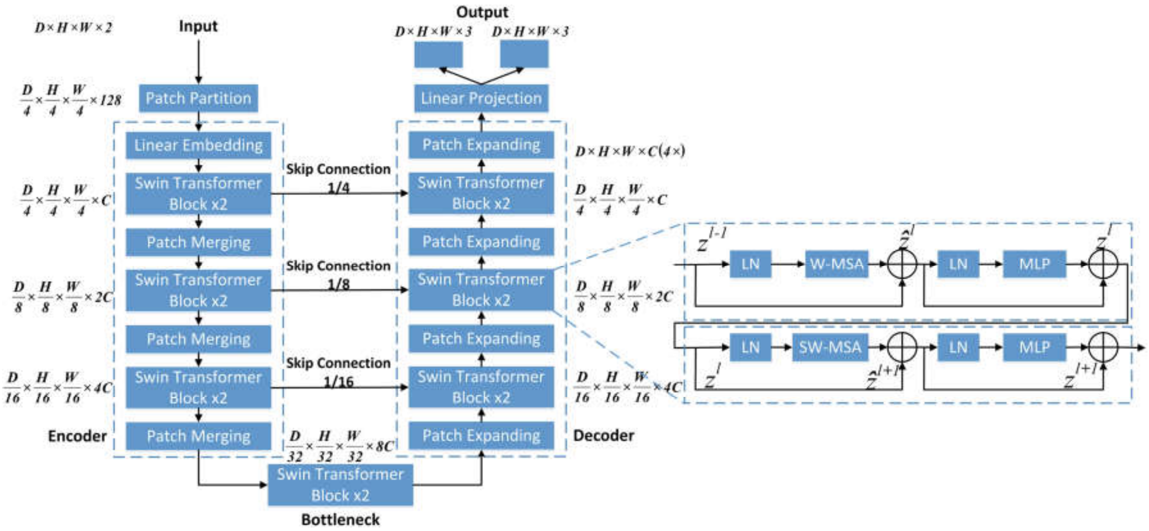
Input
$D\times H\times W\times 2$
Output
$D\times H\times W\times 3$
$D\times H\times W\times 3$
Patch Partition
$\frac{D}{4}\times\frac{H}{4}\times\frac{W}{4}\times 128$
Linear Embedding
Swin Transformer Block x2
$\frac{D}{4}\times\frac{H}{4}\times\frac{W}{4}\times C$
Patch Merging
Swin Transformer Block x2
$\frac{D}{8}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
Patch Merging
Swin Transformer Block x2
$\frac{D}{16}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
Patch Merging
Encoder
Skip Connection 1/4
Skip Connection 1/8
Skip Connection 1/16
$\frac{D}{32}\times\frac{H}{32}\times\frac{W}{32}\times 8C$
Swin Transformer Block x2
Bottleneck
Linear Projection
Patch Expanding
$D\times H\times W\times C(4\times)$
Swin Transformer Block x2
$\frac{D}{4}\times\frac{H}{4}\times\frac{W}{4}\times C$
Patch Expanding
Swin Transformer Block x2
$\frac{D}{8}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
Patch Expanding
Swin Transformer Block x2
$\frac{D}{16}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
Patch Expanding
Decoder
z^{l-1}
LN
W-MSA
$\hat{z}^{l}$
LN
MLP
z^{l}
LN
SW-MSA
z^{l}
$\hat{z}^{l+1}$
LN
MLP
z^{l+1}

Fig. 2. The architecture of Swin-Unet with Swin transformer block, which is composed of encoder, bottleneck, decoder and skip connections. Here C = 96.

2.2 Loss Function

Similar to the existing CNN-based methods [19], the optimization problem for registration is typically described as:

$$\mathcal{L}(I_1, I_2, \boldsymbol{\phi}) = \mathcal{L}_{sim}(I_1, I_2(\boldsymbol{\phi})) + \mathcal{L}_{reg}(\boldsymbol{\phi}) \tag{1}$$

$I_2(\boldsymbol{\phi})$ is I_2 warped by $\boldsymbol{\phi}$. $\mathcal{L}_{sim}$ measures image similarity between $I_2(\boldsymbol{\phi})$ and I_1, $\mathcal{L}_{reg}$ imposes regularization on $\boldsymbol{\phi}$.

Here, we set $\mathcal{L}_{sim}$ to the mean squared voxelwise difference of $I_2(\boldsymbol{\phi})$ and I_1, namely $\mathcal{L}_{sim}(I_1, I_2(\boldsymbol{\phi})) = MSE(I_1, I_2(\boldsymbol{\phi})) = \frac{1}{|\Omega|} \sum_{p\in\Omega} [I_1(p) - I_2(\boldsymbol{\phi}(p))]^2$. Specifically, our proposed similarity loss function $\mathcal{L}_{sim-pair}$ consists of two symmetric loss terms, which measures the pairwise dissimilarity between the warped M to F and warped F to M:

$$\mathcal{L}_{sim-pair}(F, M, \boldsymbol{\phi}) = \mathcal{L}_{sim}(F, M(\boldsymbol{\phi}_{MF})) + \mathcal{L}_{sim}(M, F(\boldsymbol{\phi}_{FM})) \tag{2}$$

where $\boldsymbol{\phi}_{MF}$ and $\boldsymbol{\phi}_{FM}$ are differentiable and invertible in a bidirectional fashion.

Existing learning-based methods [21] often regularize the transformation $\boldsymbol{\phi}$ with a regularization loss function, such as a L2-norm on the spatial gradients of $\boldsymbol{\phi}$. Here, we define $\mathcal{L}_{smooth}(\boldsymbol{\phi})$ as follows:

$$\mathcal{L}_{smooth}(\boldsymbol{\phi}) = \sum_{p\in\Omega} (\parallel \nabla\boldsymbol{\phi}_{MF}(p) \parallel^2 + \parallel \nabla\boldsymbol{\phi}_{FM}(p) \parallel^2) \tag{3}$$

Local Orientation Consistency Constraint. Although the smoothness of the deformation field can be controlled by the weight of the regularizer, the global regularizer may greatly degrade the registration accuracy of the model, especially when a large weight is assigned for the regularizer. In addition, these regularizers are not sufficient to guarantee a topology-preservation transformation. To address this issue, we apply a selective Jacobian determinant regularization loss $\mathcal{L}_{Jdet}(\boldsymbol{\phi})$, which imposes a local orientation consistency constraint on the estimated $\boldsymbol{\phi}$ and guarantees a topology-preservation transformation:

$$\mathcal{L}_{Jdet}(\boldsymbol{\phi}) = \frac{1}{N} \sum_{p \in \Omega} (\sigma(-(\mid J(\boldsymbol{\phi}_{MF})(p) \mid)) + \sigma(-(\mid J(\boldsymbol{\phi}_{FM})(p) \mid))) \tag{4}$$

N denotes the total number of elements in $\mid J(\boldsymbol{\phi})(p) \mid$, $\sigma(\cdot) = ReLU(\cdot)$ represents an activation function and $\mid J(\boldsymbol{\phi})(p) \mid$ denotes the Jacobian determinant of deformation field $\boldsymbol{\phi}$ at position p. The lower number of negative Jacobian determinant will lead to better diffeomorphic property of registration field [26].

Inverse Consistency Constraint. Here, we extend the objective function by adding an inverse consistency constraint $\mathcal{L}_{cons-pair}$ [1]. Inverse-consistent registration means the bidirectional deformations estimated between an image pair should share the same pathway. That is, the composition of forward and backward deformations should be identity or close to identity. Therefore, the loss $\mathcal{L}_{cons-pair}$ can be defined by:

$$\mathcal{L}_{cons-pair}(\boldsymbol{\phi}) = \frac{1}{|\Omega|} \sum_{p \in \Omega} [(\boldsymbol{\phi}_{MF} \circ \boldsymbol{\phi}_{FM})(p) - \boldsymbol{\phi}_0]^2 \tag{5}$$

During training, the composition results would be gradually approaching to identity, resulting smooth and invertible deformations. Therefore, the complete loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_{sim-pair} + \lambda_1 \mathcal{L}_{smooth} + \lambda_2 \mathcal{L}_{Jdet} + \lambda_3 \mathcal{L}_{cons-pair} \tag{6}$$

where λ_1, λ_2 and λ_3 are the weights to balance the contributions of $\mathcal{L}_{smooth}$, $\mathcal{L}_{Jdet}$ and $\mathcal{L}_{cons-pair}$, respectively.

3 Experiments

3.1 Datasets, Preprocessing and Evaluation Criteria

Datasets and Preprocessing. We validate our method on two datasets, ADNI [18] and PPMI [24], including 1961 T1-weighted brain MRI scans. And we split the datasets into 1569, 196 and 196 (8:1:1) volumes for train, validation, and test sets respectively. Specifically, we register each scan to an atlas computed using external data [17]. We focus on atlas-based registration, each input volume pair consists of the atlas (F) and a random volume (M) from the dataset. Standard preprocessing steps for structural brain MRI are performed, including skull stripping, resampling and affine spatial normalization for each scan using FreeSurfer [17], and crop the resulting images to $160 \times 192 \times 224$. Segmentation maps including 29 anatomical structures are obtained using FreeSurfer for evaluation.

Table 1. Summary of results on test set: mean Dice scores over all anatomical structures and subjects (higher is better), mean runtime, and mean percentage of locations with non-positive Jacobian (lower is better). Standard deviations are presented in parentheses. The best results are in bold.

Method	Avg. Dice	GPU sec	CPU sec	% of $\mid J(\boldsymbol{\phi})(p) \mid \leq 0$
Affine only	0.583 (0.158)	0	0	0
ANTs SyN (CC)	0.748 (0.132)	-	9054 (2021)	0.144 (0.092)
VoxelMorph (MSE) [19]	0.754 (0.140)	0.54 (0.02)	141 (1.22)	0.188 (0.082)
VoxelMorph-diff [15,16]	0.753 (0.135)	0.44 (0.01)	51 (0.22)	6.2e–6 (7.3e–5)
DeepFLASH [25]	0.761 (0.115)	0.52 (0.01)	134 (1.13)	0.175 (0.125)
SYMNet [6]	0.763 (0.113)	**0.41 (0.03)**	**48 (0.14)**	4.5e–5 (3.7e–4)
Swin-VoxelMorph (ours)	**0.775 (0.128)**	0.43 (0.01)	52 (1.12)	**2.4e–6 (2.3e–5)**

Evaluation Criteria. We expect the regions in $M(\boldsymbol{\phi})$ and F corresponding to the same anatomical structure to overlap well, and quantify the volume overlap between structures using the Dice score. The Jacobian matrix $J(\boldsymbol{\phi})(p) = \nabla\boldsymbol{\phi}(p) \in \mathbb{R}^{3\times3}$ captures the local properties of $\boldsymbol{\phi}$. We compute the numbers of all non-background voxels for which $\mid J(\boldsymbol{\phi})(p) \mid\leq 0$, where $\boldsymbol{\phi}$ is not diffeomorphic.

Implementation. Our experiments are implemented by PyTorch [2] on NVIDIA GTX 2080Ti GPUs and adopt the Adam optimizer [20] with a learning rate of 10^{-4}. We set the epochs as 1500, batch size as 1, steps of per epoch as 100. We select the model that obtain the highest Dice on the validation set and get the best results with $\lambda_1 = 0.01, \lambda_2 = 1000, \lambda_3 = 10$ which were tuned by grid search.

3.2 Results

Registration Performance. Table 1 shows average Dice and percentage of voxels with non-positive Jacobian determinant over all subjects and structures for different methods. We can see that Swin-VoxelMorph achieves the overall best performance in terms of average Dice, while producing the best diffeomorphic registration fields (less non-positive Jacobian voxels), which implies that our resulting registration fields guarantee the desirable diffeomorphic properties. Figure 3 shows the examples of moved images with overlaid boundaries of ventricles (yellow), transformation fields, and Jacobian determinant of transformations estimated by different experiments. From the red box in the upper two rows of right panel, our method captures clearer details of moved images and performs better registration than other methods, it can handle various changes in shape of structures, such as ventricles and hippocampi. The transformation fields, and estimated Jacobian determinant corresponding to the top images are shown in the lower two rows of right panel. Note that the value of colorbar indicates how volume changes in Jacobian determinant, our method produces less non-positive Jacobian voxels, which gracefully guarantees the smoothness of transformation fields without artifacts. The boxplots in Fig. 4 illustrate the Dice

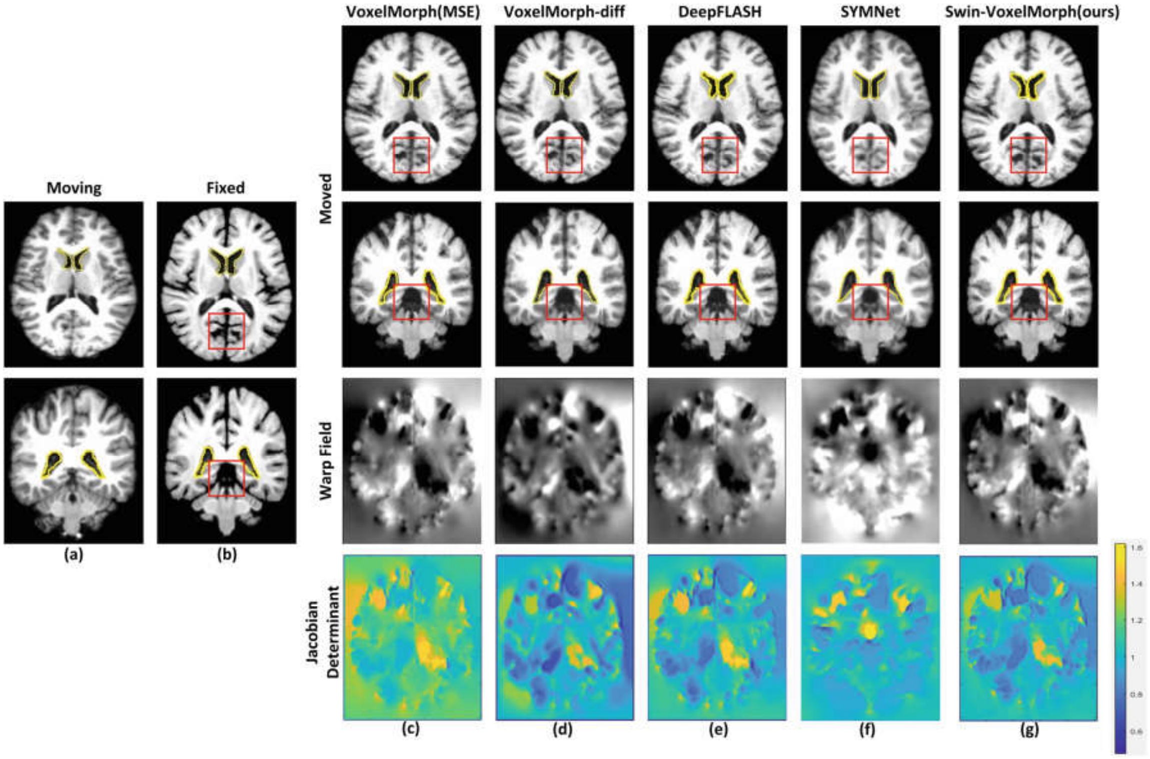

Fig. 3. Comparison of two example MR slices for different experiments. Left: (a) moving image, (b) atlas. Right: (c) (d) (e) (f) moved images, transformation fields and Jacobian determinant of transformations by VoxelMorph (MSE), VoxelMorph-diff, DeepFLASH, SYMNet and Swin-VoxelMorph.

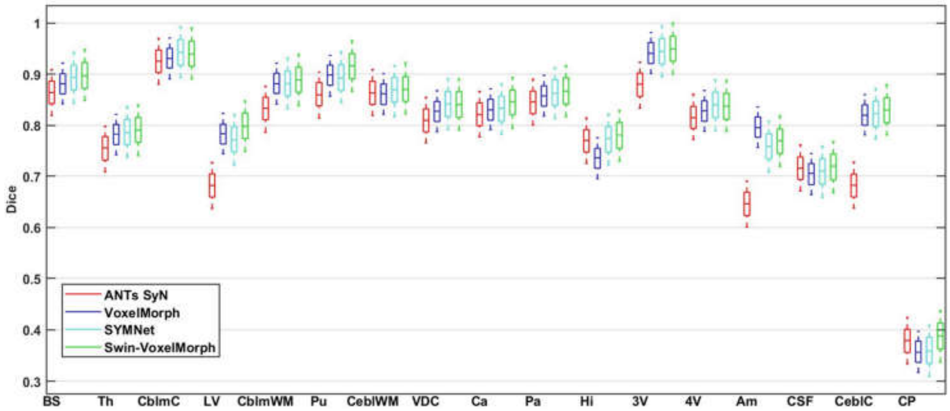

Fig. 4. Boxplots illustrating Dice scores of anatomical structures for ANTs SyN, VoxelMorph (MSE), SYMNet and Swin-VoxelMorph. Left and right brain hemispheres are combined into one structure for visualization. Brain stem (BS), thalamus (Th), cerebellum cortex (CblmC), lateral ventricle (LV), cerebellum white matter (CblmWM), putamen (Pu), cerebral white matter (CeblWM), ventral DC (VDC), caudate (Ca), pallidum (Pa), hippocampus (Hi), 3rd ventricle (3V), 4th ventricle (4V), amygdala (Am), CSF (CSF), cerebral cortex (CeblC), and choroid plexus (CP) are included.

score distribution of anatomical structures for ANTs SyN, VoxelMorph (MSE), SYMNet and our proposed method Swin-VoxelMorph. Swin-VoxelMorph performs better performance in most anatomical structures than other methods, such as Brain stem (BS), thalamus (Th), lateral ventricle (LV).

4 Conclusions

In conclusion, we propose a symmetric unsupervised learning network Swin-VoxelMorph that guarantees topology preservation and invertibility of the transformation. First, the proposed pure Transformer-based 3D U-shaped Encoder-Decoder network Swin-UNet explicitly exploit Swin Transformer to predict better registration fields for deformable medical image registration, which can provide more precise anatomical alignment. Second, our objective functions can enforce the inverse consistency of the predicted transformations. Results show that our method can outperform state-of-the-art methods in registration accuracy and the quality of diffeomorphic properties.

References

1. Shen, Z.Y., Han, X., Xu, Z.L., Niethammer, M.: Networks for joint affine and non-parametric image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4224–4233 (2019)
2. Paszke, A. et al.: Automatic differentiation in pytorch. In: NIPS-W (2017)
3. 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)
4. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
5. Zhang, J.: Inverse-consistent deep networks for unsupervised deformable image registration. arXiv preprint arXiv:1809.03443 (2018)
6. Mok, T.C.W., Chung, A.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4644–4653 (2020)
7. Miao, S., et al.: Dilated FCN for multi-agent 2D/3D medical image registration. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
8. Liao, R., et al.: An artificial agent for robust image registration. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
9. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
10. Cao, H., et al.: Swin-UNet: UNet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
11. Chen, J., et al.: TransuNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv: 2102.04306 (2021)
12. Dosovitskiy, A., et al.: An image is worth 16 x 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

13. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. IEEE Trans. Med. Imaging **32**(7), 1153–1190 (2013)
14. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. arXiv preprint arXiv: 2005.12872 (2020)
15. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A.F., Schnabel, J.A., 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
16. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. arXiv preprint arXiv: 1903.03545 (2019)
17. Fischl, B.: Freesurfer. Neuroimage **62**(2), 774–781 (2012)
18. Susanne, C., et al.: Ways toward an early diagnosis in Alzheimers disease: the Alzheimers disease neuroimaging Initiative (ADNI). Alzheimer's Dementia **1**(1), 55–66 (2005)
19. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260 (2018)
20. Diederik, P.K., Jimmy, B.: ADAM: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. 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 **99**, 1788–1800 (2019)
22. Sheikhjafari, A., Punithakumar, K.: Unsupervised deformable image registration with fully connected generative neural network. In: MIDL (2018)
23. Hou, B., Miolane, N., Khanal, B., Lee, M.: Deep pose estimation for image-based registration. In: MIDL (2018)
24. Marek, K., et al.: The Parkinson progression marker initiative (PPMI). Prog. Neurobiol. **95**(4), 629–635 (2011)
25. Wang, J., Zhang, M.M.: DeepFLASH: an efficient network for learning-based medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
26. Gu, D., et al.: Pair-wise and group-wise deformation consistency in deep registration network. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 171–180. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_17
27. Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I.: 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, pp. 204–212 (2017)
28. Luo, W.J., Li, Y.J., Urtasun, R., Zemel, R. : Understanding the effective receptive field in deep convolutional neural networks. In: Conference on Advances in Neural Information Processing Systems, vol. 29 (2016)
29. Li, S.H., Sui, X.C., Luo, X.D., Xu, X.X., Liu, Y., Goh, R.: Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511 (2021)