Symmetric Transformer-based Network for Unsupervised Image Registration

Mingrui Ma, Lei Song, Yuanbo Xu, Guixia Liu
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education,
College of Computer Science and Technology, Jilin University
mamr19,songlei@mails.jlu.edu.cn
liugx,yuanbox@jlu.edu.cn

Abstract

Medical image registration is a fundamental and critical task in medical image analysis. With the rapid development of deep learning, convolutional neural networks (CNN) have dominated the medical image registration field. Due to the disadvantage of the local receptive field of CNN, some recent registration methods have focused on using transformers for non-local registration. However, the standard Transformer has a vast number of parameters and high computational complexity, which causes Transformer can only be applied at the bottom of the registration models. As a result, only coarse information is available at the lowest resolution, limiting the contribution of Transformer in their models. To address these challenges, we propose a convolution-based efficient multi-head self-attention (CEMSA) block, which reduces the parameters of the traditional Transformer and captures local spatial context information for reducing semantic ambiguity in the attention mechanism. Based on the proposed CEMSA, we present a novel Symmetric Transformer-based model (SymTrans). SymTrans employs the Transformer blocks in the encoder and the decoder respectively to model the long-range spatial cross-image relevance. We apply SymTrans to the displacement field and diffeomorphic registration. Experimental results show that our proposed method achieves state-of-the-art performance in image registration. Our code is publicly available at https://github.com/MingR-Ma/SymTrans.

1 Introduction

Medical image registration is the fundamental and crucial branch of many medical image analysis tasks. Deformable medical image registration, a part of the medical image registration, aims to establish the dense and nonlinear correspondence between a pair of images. Traditional image methods formulate image registration as an optimization problem to search for a smooth transformation between the points in the pair of images [13, 28]. However, the traditional methods are very time-consuming and require a lot of computing resources because iterative optimization is required every time for a new image pair.

Since recently, with the rapid development of deep learning, convolutional neural networks (CNN) have been applied in many vision tasks and demonstrated the outperformance in many vision tasks [15, 23, 29]. Compared to the traditional methods in medical image registration, CNN-based methods can improve the registration performance and compute the dense transformation faster once the CNN model train is finished. However, the inherent limitation of the CNN architectures, that is, the local convolution operation (i.e., the local receptive field of CNN), makes the CNN-based methods unable to obtain the long-range spatial relations [10]. Although some approaches have been proposed to enlarge the local receptive field of CNN, they are still restricted by the kernel size of the convolution [32, 21].

The Transformer module that performs well in natural language processing tasks does not have the limitation of local receptive fields. Benefiting from the non-local receptive field capability of the Transformer, VIT [10] is the first to apply the Transformer in computer vision (CV), which regards an image as a sequence of patches (i.e., making one image into tokens), achieves the state-of-the-art image recognition results. Recently, many Transformer-based or variant Transformer-based methods have been proposed to model the CV tasks, such as Swin Transformer [17] and transU-Net [8].

In medical image registration, the size of the local receptive field of CNN itself will limit the performance of the CNN-based model to establish the correspondence between the same anatomical structures of two images, especially when the same anatomical structure is distant. Based on the Transformer studies in CV, some image registration approaches have utilized the Transformer in their methods. Vit-V-Net [7], as we know, is the first to apply the Transformer in image registration and achieves promising performance. There are also other Transformer-based image registration methods, such as DTN [31] and TransMorph [5]. However, the limited memory and a large number of parameters force them to apply the Transformer at the bottom of their networks, where the coarse feature maps are available. The lowest level resolution information limits the contribution of the Transformer.

To address these issues, we propose an encoder-decoder scheme model consisting of convolutional and Transformer blocks. We present the convolution-based efficient multi-head self-attention (CEMSA), which focuses on capturing local and long-range contextual information. Specifically, we utilize the depth-wise separable convolutional operations to capture the local contextual feature maps and compress the memory and parameters. We use our proposed patch expanding to restore the feature maps from the last CEMSA-based Transfomer encoder to build the symmetric encoder-decoder architecture. Then, the skip connections and the proposed merging operations are used to restore and fuse feature maps in the decoder. Based on these proposed modules, we build the CEMSA-Transformer-based symmetric network (SymTrans). We also introduce a variant model diff-SymTrans to obtain the diffeomorphic deformation field. Qualitative and quantitative evaluation of the experimental results demonstrates the outperformance of the proposed method in image registration.

In summary, the main contributions of this work are following:

- *CEMSA:* We propose an efficient multi-head self-attention mechanism to save memory, reduce parameters, and capture the local relevance.
- *An CEMSA-Transformer-based symmetric architecture:* We present a novelty CEMSA-Transformer-based symmetric network, SymTrans, for deformable image registration.
- *Displacement and diffeomorphic registration:* We present the two registration fashions, SymTrans, and diff-SymTrans. SymTrans yield the displacement field for registration, and diff-Trans yield the deformation field, ensuring the diffeomorphic properties.
- *State-of-the-art results:* We compare SymTrans and diff-SymTrans with three unsupervised learning-based and one wildly used traditional registration approaches. The experimental results demonstrate the state-of-the-art performance.

2 Background

2.1 Image Registration

Deformable image registration aims at establishing spatial correspondence between two images. The registration of a pair of images can be optimized by an energy function. The typical optimization problem is written as:

$$\hat{\phi} = \arg\min_{\theta} \mathbb{E}(I_m, I_f, \phi). \tag{1}$$

In this energy function, I_m and I_f denote the moving and fixed image, respectively. And ϕ denotes the deformation field, which indicates the directions and magnitudes of a spatial pixel point's transformation. $\mathbb{E}$ can be fomulated as:

$$\mathbb{E}(I_m, I_f, \phi) = \mathbb{E}_{sim}(I_m \circ \phi, I_f) + \lambda \mathbb{R}(\phi), \tag{2}$$

where $\mathbb{E}_{sim}(\cdot)$ is the similarity metric, $\circ$ is the interpolation operation, and $I_m \circ \phi$ is the warped image warped by the deformation field ϕ. The similarity function is the metric to evaluate the level of alignment between the warped moving image (i.e., $I_m \circ \phi$) and the fixed image I_f. $\mathbb{R}(\cdot)$ is a regularizer that enforces the deformation smooth. λ is the hyperparameter to balance the contributions of the similarity and the regularization.

2.2 Vision Transformer

A standard Transformer block consists of two components: multi-head self-attention (MSA) and position-wise feed forward module (FFN) [24]. Let I is an image volume defined in the 3D spatial domain $\Omega \subset \mathcal{R}^{D\times H\times W}$. To use the Transformer model the input volume, an image is first divided into N patches, then flattened to sequences of vectors $I_p \subset \mathcal{R}^{N\times P^3}$. The number of patches can be calculated by the formula $N = \frac{D\times H\times W}{P^3}$, where (*D*, *H*, *W*) is the size of the image, *P* is the size of each patch. Usually, the convolutional operation is utilized to split an image into patch embeddings without overlap [30, 10]. After getting the patch embeddings of an image, these embeddings are passed in the MSA. MSA applies the linear operation project the embeddings to the queries, keys, and values (denoted as $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$). Each linear projection set consists of *k* heads, which map the d_m dimensional input into d_k dimensional space. The input sequences to the global relations can be formulated as:

$$\text{MSA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}(\frac{\mathbf{QK^T}}{\sqrt{d_k}})\mathbf{V}. \tag{3}$$

The FFN is utilized to project the output sequence from MSA into a large scale (usually by the factor of 4) dimensional space and then project it to the sequence's original dimensional space. Thus, a Transformer block is completed.

3 Related Work

Traditional deformable image registration methods optimize the energy function formulated as Eq. 2 iteratively for each pair of images. These methods include, Demons [25], elastic model [22], and two commonly used methods SyN [1] and LDDMM [4]. These methods, as traditional methods, still face the problem of time-consuming calculations.

Unlike the traditional approaches, CNN-based methods learn the parameters of their model on the training dataset to predict the deformation field between a pair of unseen images. Therefore, CNN-based methods compute the deformation field usually less than a second (after training). The CNN-based methods can be categorized as supervised and unsupervised. The supervised methods require the ground-truth information in the dataset, while the ground-truth deformation fields are hard to obtain [20, 11]. Comparing with the supervised registration methods, unsupervised methods are not limited to the ground-truth information. According to the output of a methods, the registration methods can be divided into two categories: the displacement field registration and the diffeomorphic registration. The diffeomorphic methods compute the diffeomorphic deformation field to guarantee the desirable diffeomorphic property [19, 9, 26, 16]. The displacement field methods output the deformation field directly from their CNN model, which directly use the deformation field to warp the moving image toward the fixed image [3, 7]. Some recent studies employ the Transformer at the bottom of their network to overcome CNN's local receptive field shortcoming [7, 31]. The reason for placing the Transformer at the bottom of their networks is that the memory and computational complexity significantly increase with the higher resolution level. Motivated by the latest researches [27, 30], we propose the CEMSA. Based on CEMSA, we build the CEMSA-Transformer-based symmetric network consisting of a total of ten Transformer blocks in 1/4, 1/8, and 1/16 resolution levels to enhance the contribution of transformers.

4 Methods

Let a pair of images be defined in the spatial domain $\Omega \subset \mathcal{R}^n, (n = 3)$. Fig. 1 illustrates the overall architectures of deformable image registration in this paper. Briefly, the moving and fixed images (respectively denoted as I_f and I_m) first input the proposed Transformer-based network, and then the network outputs the deformation field. Finally, the spatial transformation network [14] is utilized to warp the moving image toward a fixed image via the deformation field. $\mathcal{L}_{sim}$ is the similarity loss function to evaluate the similarity between warped and fixed images. $\mathcal{L}_{reg}$ is the regularization to enforce the magnitude of the deformation field.

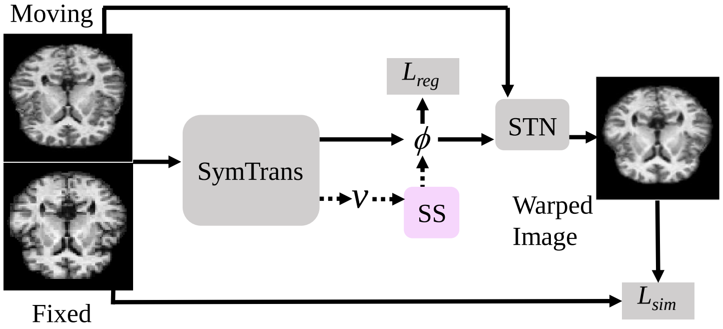

Figure 1: Overview of the proposed method for deformable image registration. The pink block named SS represents the scaling-and-squaring module. STN represents the spatial transform network. The dotted line indicates the workflow for diffeomorphic registration.

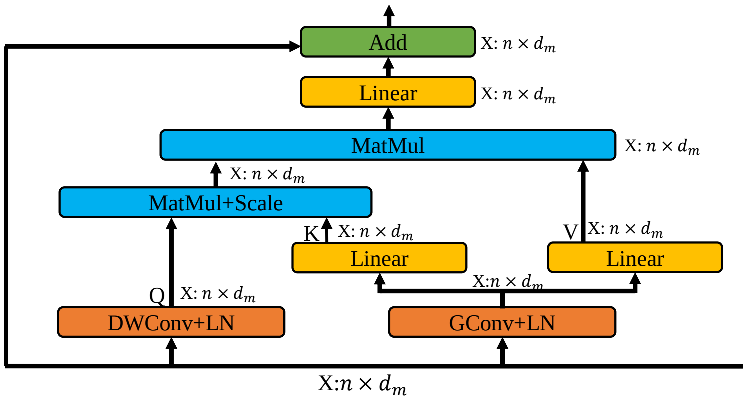

Figure 2: The proposed CEMSA block.

4.1 Efficient Transformer Block

The standard Transformer usually takes up a lot of memory because Transformer has a large number of parameters, especially when applied in the 3D image tasks. To build the Transformer blocks symmetrically both in the encoder and decoder, we present a novel convolution-based efficient multi-head self-attention (CEMSA) for the Transformer block in this paper. The proposed CEMSA is shown in Fig. 2. Compared with the standard Transformer, we employ the depth-wise separable and grouped convolution in the proposed CEMSA, which can further capture local spatial context, and reduce the semantic ambiguity and the computation costs. Each token input for attention function of **Q**, **K**, **V** can be summarily formulated as:

$$x^{q,k,v} = \text{Flatten}(\text{Conv3D}(\text{Reshape}(x), s)), \tag{4}$$

where x is the input tokens to the CEMSA. DWConv is the depth-wise convolutional operation with the kernel size of s. GConv is the grouped convolutional operation with the number of the groups of the input's dimensions. After the DWConv and GConv, the LN (layer normalization) is applied. Then, two linear projection sets are utilized to obtain **K** and **V**. After that, we adopt Eq. 3 to compute the attention function on **Q**, **K**, **V**. We use different s for the depth-wise convolutional operation at 1/4, 1/8, 1/16 resolution levels. Then, we take advantage of the standard FFN to project the output of CEMSA. Thus, an CEMSA-based Transformer block is constructed.

It is worth noting that compared with the Standard and efficient MSA block mentioned and proposed in [10, 30], we remove the position embedding to reduce the parameters further. [27] illustrates that the Transformer with the convolutional projection does not require position embedding because the convolutional projection represents the continuous positional information between tokens. To weaken the affection of eliminating the position embedding, and fully guaranteeing the positional information of each token, we use the single DWConv to compute the attention function on Q to get the spatial positional information. Compared to the existing efficient MSA approach [30], we do not

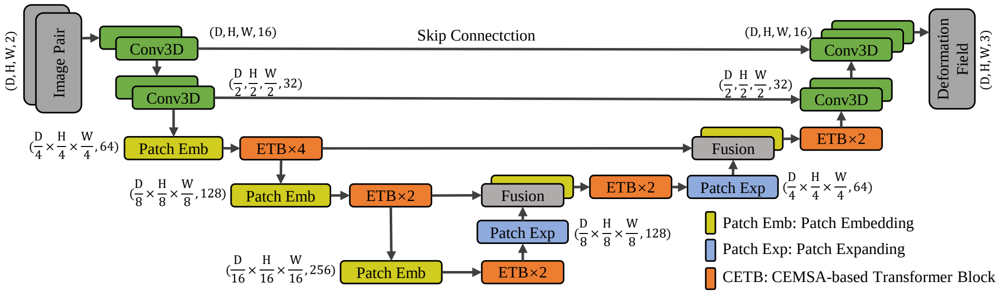

Figure 3: The propsed symmetric Transformer-based netowrk.

use the linear projection after the DWConv to maintain the positional information of each token. The role of the GConv operation is to reduce the parameters before the linear projections. According to the set number of groups g, GConv can reduce the number of parameters to $1/g$. We adopt Eq. 3 to compute the attention on Q, K, and V. As a result, the proposed CEMSA takes into account the spatial positional relations and the information of tokens at different positions.

4.2 Symmetric Transformer-based Network

Using the proposed CEMSA-based Transformer, we can build the CEMSA-based Transformer blocks (SymTrans) both in encoder and decoder. The proposed SymTrans is shown in Fig. 3. The SymTrans is a U-shaped model like U-Net, consisting of 2 CNN-based encoding-decoding layers and 3 Transformer-based encoding-decoding layers. Each of the Transformer-based encoding-decoding blocks requires a sequence input. We utilize the convolutional operations to perform the *Patch Embedding* operations before each Transformer in the encoder, with the stride of 2, kernel size of 3 (i.e., the patch size), to obtain the patch sequences with overlap. Before the feature maps input the next level Transformer block in the decoder, we utilize the *Patch Expanding* operations to enlarge the feature maps.

In detail, the *Patch Expanding* operations consists of two linear projections, first expanding the size of feature maps by the factor of 2^3, then expanding the feature maps dimension by the factor of 2. In the gap of SymTrans, the skip connections are used to concatenate the output feature maps from the Transformer in the encoder and the expanded feature maps from the decoder. Then, the *Fusion* operations are utilized to reshape these two sequence feature maps to image form, then fuse them using the convolutional operations. At the half and original size resolution level, we utilize the convolutional blocks with the kernel size of 3, the stride of 1 (to the same resolution level), the stride of 2 (to the next resolution level) for encoding and decoding.

4.3 Registration and Learning

4.3.1 Registration

In this paper, we apply the SymTrans to displacement field registration and diffeomorphic registration. As shown in Fig. 1, the deformation field can be generated in two ways to register a pair of images: the solid line following the SymTrans indicates the displacement field registration, the dotted line following the SymTrans indicates the diffeomorphic registration. The diffeomorphic branch ensures the diffeomorphism in registration. The diffeomorphism is a continuous, invertible, and one-to-one mapping. To achieve that, we follow [9, 19] and use the stationary velocity field with the efficient scaling-and-squaring approach to obtain the diffeomorphic deformation field. In the scaling-and-squaring approach, the deformation field is represented as a Lie algebra member that is exponentiated to generate the deformation field at time 1, which is a member of the Lie group, can be written as $\phi^{(1)} = \exp(v)$. Starting from the initial deformation field at time 0, i.e., the output velo city field from the SymTrans, can be formulated as:

$$\phi^{(1/2^T)} = p + \frac{v(p)}{2^T}, \tag{5}$$

where *p* is the map of spatial locations. The recurrence to obtain the deformation field at time 1 can be written as:

$$\phi^{(1/2^{t-1})} = \phi^{(1/2t)} \circ \phi^{(1/2t)}. \tag{6}$$

Hence, the time 1 deformation field $\phi^{(1)} = \phi^{(1/2)} \circ \phi^{(1/2)}$ is obtained.

4.3.2 Learning

The proposed SymTrans is optimized in an unsupervised manner by evaluating the similarity between aligned and fixed image. As shown in Fig. 1, given a image pair (I_m,I_f), the Symtrans estimates the deformation field ϕ. Then, the STN warps I_f to obtain the warped image $\hat{I}_m$ (denoted as $\hat{I}_m = I_m \circ \phi$). We apply the L_2 loss both on the registration similarity and smooth regularization. The loss function is defined as Eq. 2 and formulated as $L = L_{sim}(I_f, \hat{I}_m) + \lambda L_{reg}(\nabla\phi)$. We optimize the parameters of SymTrans by minimizing this loss function.

5 Experiments

5.1 Dataset and Metrics

We demonstrate the proposed method on the task of brain MRI registration. We use the publicly available dataset OASIS, consisting of 425 T1-weighted brain MRI scans [18], and 270 scans are selected in this dataset for our experiment. We first resample each scan to $256 \times 256 \times 256$ with the isotropic voxels size of $1mm \times 1mm \times 1mm$. Then, we conduct the standard preprocessing operation to normalize, affine transformation, and strip the skull using FreeSurfer [12]. The segmentation maps of each scan viewed as ground truth for evaluation also is obtained through FreeSurfer. Each scan is cropped to $160 \times 192 \times 224$, then resampled to $96 \times 112 \times 96$. The dataset is split into 200, 34, and 36 scans for train, validation, and test sets, respectively. We sequentially, without repetition, combine two scans in the training set to obtain 39,800 permutations of image pairs. These scan pairs are used for training our proposed, and the baseline approaches. We conduct the basis atlas-based registration on the test set. Six and thirty scans are selected randomly as the atlas and moving images, respectively.

Baseline methods and proposed methods are evaluated using the Dice similarity coefficients (DSC), which calculates the overlapping between the ground truth segmentation maps and the warped moving image corresponding segmentation maps. We count the negative Jacobian determinant $|J(\phi)| \leq 0$ to denote the number of folding. $|J(\phi)| \leq 0$ relates where the voxels lose topology preservation and the violate the diffeomorphic property when transformed via the deformation field.

5.2 Baseline Methods

We compare the proposed method SymTrans with five approaches, including one traditional and four deep-learning methods. The symmetric image normalization registration method (SyN) is a traditional iterative method to compute the deformation field [1]. We use the SyN implementation in the ANTs [2] toolbox and set the iteration to [100,100,100]. The deep learning baseline methods, including the CNN-based VoxelMorph [3], the CNN-based SYMNet [19], the Transformer-based ViT-V-Net [7] and the Swin-Transformer-based TransMorph [6]. We use the publicly available implementation of these four deep learning methods. We train the VoxelMorph, SYMNet, Vit-V-Net and TransMorph with the setting of their suggested hyperparameters, on the same data set splitting, respectively.

5.3 Implementation Details

The proposed framework is implemented by using the PyTorch. The STN in our method is the same as the one utilized in VoxelMorph , Vit-V-Net, and TransMorph. We set the regularizing parameter λ to 0.02. We employ the Adam optimizer to optimize the parameters of the proposed network, with a learning rate of 1e-4, on an NVIDIA RTX3080 10 GB GPU. The maximum iterations of training for the deep-learning approaches are 300k.

The detailed configures of the proposed CEMSA-based Transformer during training is following: $s = \{24, 16, 12\}$ at 1/4, 1/8, 1/16 resolution stages; the number of heads is $\{2, 4, 8\}$ at each resolution

Method	DSC	$\lvert J(\phi)\rvert \leq 0$
Affine	0.520 (0.058)	-
SyN	0.662 (0.038)	40.683 (78.042)
VoxelMorph	0.726 (0.031)	1453.778 (624.714)
SYMNet	0.719 (0.025)	1205.789 (365.011)
Vit-V-Net	0.730 (0.031)	1563.088 (631.037)
TransMorph	0.742 (0.027)	1631.978 (574.568)
SymTrans	**0.747 (0.026)**	1581.033 (587.560)
diff-SymTrans	**0.742 (0.025)**	**2.033 (9.942)**

Table 1: Qualitative comparison between our frameworks and baseline methods. DSC higher is better, and $\lvert J(\phi)\rvert \leq 0$ lower is better. Standard deviations are in bracket.

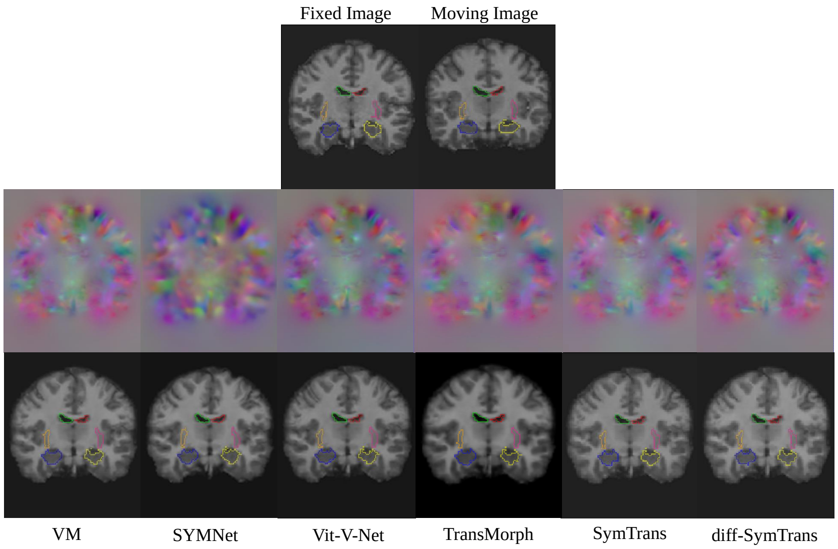

Figure 4: The atlas-based registration of lateral-ventricle, thalamus, and hippocampus by the VoxelMorph, SYMNet, Vit-V-Net, TransMorph, and the proposed SymTrans and diff-SymTrans.

stage; the patch size is $\{3, 3, 3\}$; the number of the grouped convolution's groups is equal to the input embedding dimension.

5.4 Results

5.4.1 Registration Accuracy

Fig. 4 shows the registration results of a pair of images. The boundaries of three segmentation maps are marked in the sampled slices to observe the deformation of each anatomical structure. We quantitatively evaluate the accuracy of the baseline methods and the proposed SymTrans using the DSC metric. The non-positive Jacobian determinants are utilized to assess the number of folding. Table 1 shows the results of different methods on the same test set. The proposed SymTrans, applied to displacement field registration, produces the highest average DSC than the baseline methods. The diffeomorphic registration using SymTrans (denoted as diff-SymTrans) still gives the higher average DSC than baseline methods and decreases the average number of folding much lower, which guarantees the topology of the original moving image. Besides, the lower standard deviations of the SymTrans and diff-SymTrans show strong stability of the proposed SymTrans.

To demonstrate the alignment results of each anatomical structure, we report the DSCs of 35 anatomical structures in Fig. 5.The abbreviations in Fig. 5 are: Brainstem (BS), thalamus (Th), cerebellum cortex (CblmC), lateral ventricle (LV), putamen (Pu), pallidum (Pa), cerebral white matter

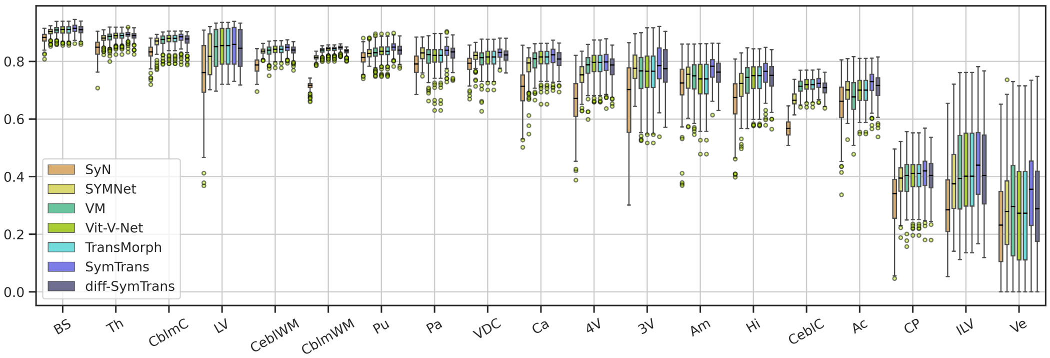

Figure 5: A boxplot illustrating the DSC of each anatomical structure segmentation for SyN, VoxelMorph, SYMNet, Vit-V-Net, and ours. We averaged the Dice values of the left and right brain hemispheres and combined them into one structure for visualization.

Method	Trans. L.	Params (M)	FLOPs (G)
VoxelMorph	-	0.29	59.82
SYMNet	-	1.12	44.51
Vit-V-Net	1/16	31.50	65.77
TransMorph	1/4	46.69	112.75
SymTrans	1/4	16.05	63.53

Table 2: The parameters and FLOPs comparison of different methods for registration. Input image size is 96 × 112 × 96 by default. Trans. L.: The starting deployment location of the Transformer.

(CeblWM), ventral DC (VDC), caudate (Ca), Amygdala (Am) hippocampus (Hi), 3rd ventricle (3V), 4th ventricle (4V), amygdala(Am), CSF (CSF), cerebral cortex (CeblC), inf-lateral ventricle (ILV), Vessel (Ve) and choroid plexus (CP). As we can see, the proposed SymTrans outperforms the compared registration approaches on all of the 19 combined structures. The diff-SymTrans yields better results than all baseline methods except TransMorph and SymTrans, while producing minimal folding. To sum up, the proposed symmetric Transformer model based on the CEMSA achieves the best results.

5.4.2 Computational Complexity

To illustrate the effectiveness of the proposed CEMSA, we compare its parameters and the FLOPs with the baseline approaches. Table 3 shows the FLOPs and the parameters of each method. The CNN-based networks, both VoxelMorph and SYMNet, have fewer parameters and FLOPs than Transformer-based models because Transformer-based models have many linear operations, which enlarge the parameters and FLOPs scale. Among these three Transformer-based methods, our SymTrans achieves the lowest FLOPs. Compared with Vit-V-Net and TransMorph, the parameters of the SymTrans are much fewer, and the FLOPs are fewer than theirs. Specifically, Vit-V-Net employs 12 Transformer blocks at the bottom of their model, each block containing 1.76M parameters. TransMorph employs the Swin-Transformer blocks at the 1/4 resolution stage, which model the input embedding patches with the dimensions of 96. In the SymTrans encoder, the depth of the CEMSA-based Transformer at each resolution stage is equal to the depth of the Swin-Tranformer in TransMorph.

In general, the parameters are gained while the size of the input token raises. Symtrans applies the CEMSA-based Transformer blocks at 1/4, 1/8, and 1/16 resolution levels in the framework. Even applying the CEMSA-based Transformer blocks at so manyresolution stages, SymTrans has about 49% fewer parameters than Vit-V-Net, and 67% fewer parameters than TransMorph. In practice, the GPU memory occupied during training is about 3 GB with a batch size of 1 and an input image size of 96 × 112 × 96 on our server. Vit-V-Net and Transmorph occupy about 6 GB and 7 GB of GPU memory with the input padded image size of 96 × 128 × 96. Statistical results of parameters and

Method	E-SymTrans	D-SymTrans	B-SymTrans	SymTrans
DSC	0.734 (0.028)	0.717 (0.033)	0.714 (0.034)	0.740 (0.027)

Table 3: Comparison of placing the CEMSA-based Transformer in different branch of the proposed network. Standard deviations are in bracket.

FLOPs indicate that the proposed CEMSA is feasible to reduce parameters, which provides a basis to apply the Transformer at the high-resolution levels.

5.4.3 Ablation Studies

We investigate the performance when the CEMSA-based Transformer is applied at different locations in the network to demonstrate that the symmetric framework is effective. The original Symtrans and all the ablation are utilized to perform the displacement field registration. We train the ablation variants for 100k iterations. Then, we find the best weights on the validation set and test these variants on the test set.

Table 3 reports the DSC results of three variant SymTrans. E-SymTrans contains the CEMSA-based Transformer blocks in the encoder and replaced the CEMSA-based Transformer blocks with the convolutional blocks in the decoder. D-SymTrans indicates that only the CEMSA-based Transformer blocks are utilized in the decoder, and the rest, as shown in Fig. 3, are convolutional blocks. *Patch Embedding* and the *Fusion* blocks in these two ablations are replaced with the basis convolutional blocks. B-SymTrans is the CNN-based architecture that applies 10 CEMSA-based Transformer blocks at the bottom. Each convolution block is followed by a LeakyReLU activation to construct a Conv block. The depths of Conv blocks are the same as the depths of the replaced CEMSA-based Transformer. *Patch Expanding* blocks are replaced with the deconvolutional operation. The structures form of E-SymTrans and D-SymTrans correspond to the structures form of TransMorph and Vit-V-Net. We observe that the original SymTrans achieves the best performance. The results of these ablation variants identify that employing the CEMSA-based Transformer at the high-resolution levels of the network and applying them symmetrically as encoder and decoder enhance the registration accuracy. That demonstrates that modeling high-resolution feature maps with the symmetric architecture can facilitate the model to recognize meaningful semantic correspondences to anatomical structures.

6 Conclusion

This paper proposes an CEMSA mechanism to capture local spatial context, reduce semantic ambiguity and parameters. Based on the proposed CEMSA, we build the Symtrans for deformable image registration, which takes advantage of the long-range spatial relevance for feature enhancements. The Transformer blocks based on CEMSA are not only applied at the bottom but also at the higher-resolution levels both in encoder and decoder. The qualitative and quantitative evaluations demonstrate that the SymTrans promotes the semantically meaningful correspondence of anatomical structures and provide the state-of-the-art registration performance. Furthermore, the ablation studies illustrate the impact on performance when the Transformer is applied on different components (i.e., encoder and decoder) of the model, which indicates the effectiveness of symmetric scheme and the importance of building transformers at the high-resolution levels.

References

[1] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008.

[2] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.

[3] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. An unsupervised learning model for deformable medical image registration. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[4] M. Faisal Beg, Michael I. Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

[5] Junyu Chen, Yong Du, Yufan He, William P Segars, Ye Li, and Eirc C Frey. Transmorph: Transformer for unsupervised medical image registration. *arXiv preprint arXiv:2111.10480*, 2021.

[6] Junyu Chen, Yong Du, Yufan He, William P Segars, Ye Li, and Eirc C Frey. Transmorph: Transformer for unsupervised medical image registration. *arXiv preprint arXiv:2111.10480*, 2021.

[7] Junyu Chen, Yufan He, Eric C. Frey, Ye Li, and Yong Du. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration, 2021.

[8] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[9] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019.

[10] 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.

[11] Jingfan Fan, Xiaohuan Cao, Zhong Xue, Pew-Thian Yap, and Dinggang Shen. Adversarial similarity network for evaluating image alignment in deep learning based registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 739–746. Springer, 2018.

[12] Bruce Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012. 20 YEARS OF fMRI.

[13] Thomas Gerig, Kamal Shahim, Mauricio Reyes, Thomas Vetter, and Marcel Lüthi. Spatially varying registration using gaussian processes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 413–420. Springer, 2014.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

[15] Xue Lin, Qi Zou, and Xixia Xu. Action-guided attention mining and relation reasoning network for human-object interaction detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1104–1110, 2021.

[16] Risheng Liu, Zi Li, Yuxi Zhang, Xin Fan, and Zhongxuan Luo. Bi-level probabilistic feature learning for deformable image registration. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 723–730, 2021.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[18] 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*, 19(9):1498–1507, 09 2007.

[19] Tcw Mok and Acs Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] 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*, pages 266–274. Springer, 2017.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[22] Dinggang Shen and C. Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*, 21(11):1421–1439, 2002.

[23] Chunwei Tian, Yong Xu, Wangmeng Zuo, Bob Zhang, Lunke Fei, and Chia-Wen Lin. Coarse-to-fine cnn for image super-resolution. *IEEE Transactions on Multimedia*, 23:1489–1502, 2021.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[25] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[26] Jian Wang and Miaomiao Zhang. Deepflash: An efficient network for learning-based medical image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[27] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[28] Yue Wu, Wenping Ma, Maoguo Gong, Linzhi Su, and Licheng Jiao. A novel point-matching algorithm based on fast sample consensus for image registration. *IEEE Geoscience and Remote Sensing Letters*, 12(1):43–47, 2014.

[29] Gege Zhang, Qinghua Ma, Licheng Jiao, Fang Liu, and Qigong Sun. Attan: Attention adversarial networks for 3d point cloud semantic segmentation. In *Proceedings of the Twenty-Ninth International Conference*

on International Joint Conferences on Artificial Intelligence, pages 789–796, 2021.

[30] Qinglong Zhang and Yubin Yang. Rest: An efficient transformer for visual recognition. *arXiv preprint arXiv:2105.13677*, 2021.

[31] Yungeng Zhang, Yuru Pei, and Hongbin Zha. Learning dual transformer network for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–138. Springer, 2021.

[32] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.