Knowledge-Based Systems 257 (2022) 109959

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Symmetric transformer-based network for unsupervised image registration

Mingrui Ma, Yuanbo Xu, Lei Song, Guixia Liu [*]

Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, 130000, PR China

ARTICLE INFO

Article history:
Received 14 May 2022
Received in revised form 27 September 2022
Accepted 27 September 2022
Available online 1 October 2022

Keywords:
Medical image registration
Deep learning
Vision transformer
Symmetric transformer-based architecture

ABSTRACT

Medical image registration is a fundamental and critical task in medical image analysis. With the rapid development of deep learning, convolutional neural networks (CNNs) have dominated the medical image registration field. Due to the disadvantage of the local receptive field of CNNs, some recent registration methods have focused on using transformers for nonlocal registration. However, the standard transformer has a vast number of parameters and high computational complexity, which means that it can only be applied at the bottom of registration models. As a result, only coarse information is available at the lowest resolution, limiting the contribution of the transformer in these models. To address these challenges, we propose a convolution-based efficient multihead self-attention (CEMSA) block, which reduces the number of parameters of the traditional transformer and captures local spatial context information to reduce 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 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 step in many medical image analysis tasks. Deformable medical image registration, which is a type of image registration, aims to establish a 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 a pair of images [1,2]. However, the traditional methods are very time-consuming and require a substantial amount of computing resources because iterative optimization is required every time for a new image pair.

Recently, with the rapid development of deep learning, convolutional neural networks (CNNs) have been applied and have demonstrated superior performance in many vision tasks [3–5]. Compared with 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 training is finished. However, the inherent limitation of the CNN architectures, that is, the local convolution operation (i.e., the local receptive field of the CNN), makes the CNN-based methods obtain the spatial relations in the range of the kernel size. Although various approaches have been proposed for enlarging the local receptive fields of CNNs, they are still restricted by the small kernel size of the convolution [6,7].

The transformer module that performs well in natural language processing tasks does not have the limitation of local receptive fields. Benefiting from the nonlocal receptive field capability of the transformer, VIT [8] is the first to apply the transformer in computer vision (CV), which regards an image as a sequence of patches (i.e., transforming the image into tokens) and achieves state-of-the-art image recognition results. Recently, many transformer-based or variant transformer-based methods have been proposed for modeling CV tasks, such as the Swin transformer [9] and transU-Net [10].

In medical image registration, CNNs give more attention to the information inside the receptive field; this may limit the performance of CNN-based models in establishing correspondences between the same anatomical structures in two images, especially when the same anatomical structures are distant [11,12]. Based on transformer studies in CV, some image registration approaches have utilized the transformer in their methods. Vit-V-Net [13], is the first to apply the transformer in image registration and achieves promising performance. There are also other

[*] Corresponding author.
E-mail addresses: mamr19@mails.jlu.edu.cn (M. Ma), yuanbox@jlu.edu.cn (Y. Xu), songlei20@mails.jlu.edu.cn (L. Song), liugx@jlu.edu.cn (G. Liu).

https://doi.org/10.1016/j.knosys.2022.109959
0950-7051/© 2022 Elsevier B.V. All rights reserved.

transformer-based image registration methods, such as DTN [12] and TransMorph [14]. However, since 3D volumetric data and the transformer consume considerable GPU memory, scholars need to make some compromises in their models [11–13,15]. In the field of medical image segmentation, [15] proposes AFTer-Unet, which uses a general 2D encoder and decoder, along with a proposed axial fusion transformer encoder at the bottom of their model. The transformers of AFTer-Unet model the information between axial slices at the bottom, thereby saving GPU memory in modeling of 3D information. In medical image registration, [12,13] also apply transformers at the bottom of their models. However, this compromise of applying the transformer at the bottom of the model causes the transformer to obtain feature information at only the lowest resolution level, which limits the performance of the transformer. If the transformer can be used at a higher resolution level (i.e., if the transformer can obtain more information), the contribution of the transformer in these models can be further improved, thereby improving the performance of the models.

To address these issues, we propose an encoder–decoder scheme model consisting of convolutional and transformer blocks. We present convolution-based efficient multihead self-attention (CEMSA), which focuses on capturing local and long-range contextual information. Specifically, we utilize depthwise separable convolution operations to capture the local contextual feature maps and compress the memory and parameters. We use our proposed patch expansion to restore the feature maps from the last CEMSA-based transformer encoder to build a symmetric encoder–decoder architecture. Then, skip connections and 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, namely, diff-SymTrans, for obtaining a diffeomorphic deformation field. Qualitative and quantitative evaluations of the experimental results demonstrate the superior performance of the proposed method in image registration.

In summary, the main contributions of this work are as follows:

- *CEMSA:* We propose an efficient multihead self-attention mechanism that saves memory, reduces the number of parameters, and captures the local relevance.
- *A CEMSA transformer-based symmetric architecture:* We present a novel CEMSA transformer-based symmetric network, SymTrans, for deformable image registration.
- *Displacement and diffeomorphic registration:* We present the two registration models, namely, SymTrans and diff-SymTrans. SymTrans yields the displacement field for registration, and diff-Trans yields the deformation field, ensuring the diffeomorphicity properties.
- *State-of-the-art results:* We compare SymTrans and diff-SymTrans with three unsupervised learning-based registration methods and one widely used traditional registration approach. The experimental results demonstrate the state-of-the-art performances of the proposed models.

2. Background

2.1. Image registration

Deformable image registration aims at to establish a spatial correspondence between two images. The registration of a pair of images can be optimized by an energy function. The typical optimization problem is expressed 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 images, respectively, and ϕ denotes the deformation field, which indicates the directions and magnitudes of a spatial pixel point's transformation. $\mathbb{E}$ can be formulated 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 image warped by the deformation field ϕ. The similarity function is the metric used to evaluate the level of alignment between the warped moving image ($I_m \circ \phi$) and the fixed image I_f. $\mathbb{R}(\cdot)$ is a regularizer that enforces the smoothness of the deformation field. λ is a hyperparameter that balances the contributions of the similarity and the regularization.

2.2. Vision transformer

A standard transformer block consists of two components: multihead self-attention (MSA) and a positionwise feed-forward module (FFN) [16]. Let I be 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 and then flattened into 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 and P is the size of each patch. Usually, a convolution operation is utilized to split an image into patch embeddings without overlap [8,17]. After obtaining the patch embeddings of an image, these embeddings are passed in the MSA. The MSA applies a linear operation to project the embeddings to the queries, keys, and values (denoted as $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$). 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

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

The FFN is utilized to project the output sequence from the MSA into a higher-dimensional (usually by a factor of 4) 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 [18], the elastic model [19], and two commonly used methods, namely, SyN [20] and LDDMM [21]. 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 in usually less than a second (after training). CNN-based methods can be categorized as supervised and unsupervised. Supervised methods require the ground-truth information in the dataset, while the ground-truth deformation fields are difficult to obtain [22,23]. Compared with supervised registration methods, unsupervised methods are not limited to the ground-truth information. According to the output of these registration methods, they can be divided into two categories: displacement field registration and diffeomorphic registration. Diffeomorphic methods compute the diffeomorphic deformation field to guarantee the desirable diffeomorphicity property [24–27]. Displacement field methods output the deformation field directly from the CNN model, which directly uses the deformation field to warp the moving image toward the fixed image [13,28]. Some recent studies employ the

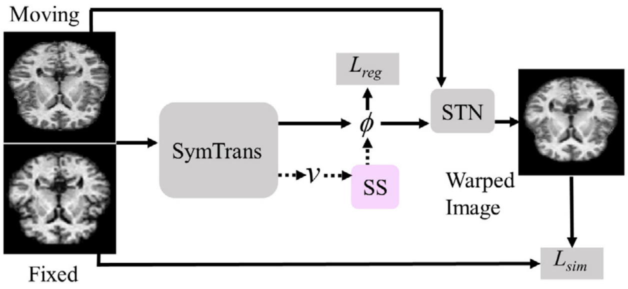

Fig. 1. Overview of the proposed method for deformable image registration. The pink block named SS represents the scaling-and-squaring module. STN represents the spatial transform network. The dotted line indicates the workflow for diffeomorphic registration. v is the velocity field for diffeomorphic branch. ϕ is the deformation field for registration.

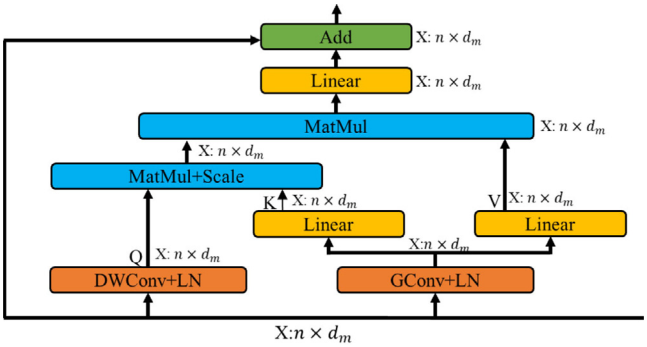

Fig. 2. The proposed CEMSA block.

transformer at the bottom of the network, and these methods achieve good registration performance [12,13]. The reason for placing the transformer at the bottom of the network is that the memory and computational complexity significantly increasing with the higher resolution level. Motivated by the latest research [17,29], we propose CEMSA. Based on CEMSA, we build a CEMSA transformer-based symmetric network consisting of a total of ten transformer blocks at 1/4, 1/8, and 1/16 resolution levels to enhance the contribution of the transformers.

4. Methods

We define a pair of images 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 (denoted I_m and I_f, respectively) are first input into the proposed transformer-based network, and then the network outputs the deformation field. Finally, the spatial transformation network [30] is utilized to warp the moving image toward a fixed image via the deformation field. The similarity loss function $\mathcal{L}_{sim}$ is used to evaluate the similarity between the warped and fixed images. The regularization function $\mathcal{L}_{reg}$ is used to enforce the magnitude of the deformation field.

4.1. Convolution-based efficient transformer block

The standard transformer usually uses considerable memory because it has a large number of parameters, especially when applied to 3D image tasks. To build the transformer blocks symmetrically in both the encoder and decoder, we present a novel convolution-based efficient multihead self-attention (CEMSA) for the transformer block in this paper. The proposed CEMSA is illustrated in Fig. 2. Compared with the standard transformer, we employ the depthwise separable and grouped convolution in the proposed CEMSA, which can further capture local spatial context and reduce the semantic ambiguity and the computational costs. Each token input for the attention function of **Q**, **K**, and **V** can be summarily formulated as

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

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

4.2. Symmetric transformer-based network architecture

Using the proposed CEMSA-based transformer, we build CEMSA-based transformer blocks (SymTrans) in both the encoder and decoder. The proposed SymTrans is illustrated in Fig. 3. SymTrans is a U-shaped model similar to 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.

The convolutional blocks in the 1/1 and 1/2 levels, which consist of two sequential convolution operations and are followed by an instance normalization (IN) layer, extract features from the input concatenated images. At the 1/4 resolution level, we utilize the convolution operations with a stride of 2 and a kernel size of 3 (i.e., the patch size) to perform the *patch embedding* operations before each transformer in the encoder to obtain flattened patch sequences (tokens) with overlap. Then, these sequences are input into the specified depth CETB. Until the bottom of SymTrans, before the feature maps are input into 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 consist of two linear projections followed by an LN operation. First, *patch expanding* expands the dimension of each feature map by 2^3. Then, it reshapes the feature maps into twice the original shape. Finally, it reduces the dimension of each feature map by half via projection. In the SymTrans gap, skip connections are used to concatenate the output feature maps from the transformer in the encoder and the expanded feature maps from the decoder. *Fusion* operations are utilized to reshape these two sequence feature maps into image form and then fuse them using convolution operations with a kernel size of 3 and a stride of 1.

4.3. Learning for SymTrans

4.3.1. Displacement registration and diffeomorphic registration

In this paper, we apply 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 SymTrans indicates displacement field registration, and the dotted line following SymTrans indicates diffeomorphic registration. The diffeomorphic branch ensures the diffeomorphism in registration. A diffeomorphism is a continuous, invertible, and one-to-one mapping. To achieve that, we follow [24,25] and use a stationary velocity field with the efficient scaling-and-squaring approach to obtain the diffeomorphic deformation field. In the scaling-and-squaring approach, the

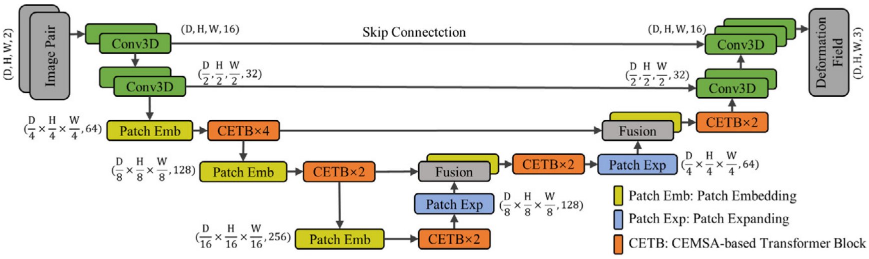

Fig. 3. The proposed symmetric transformer-based network.

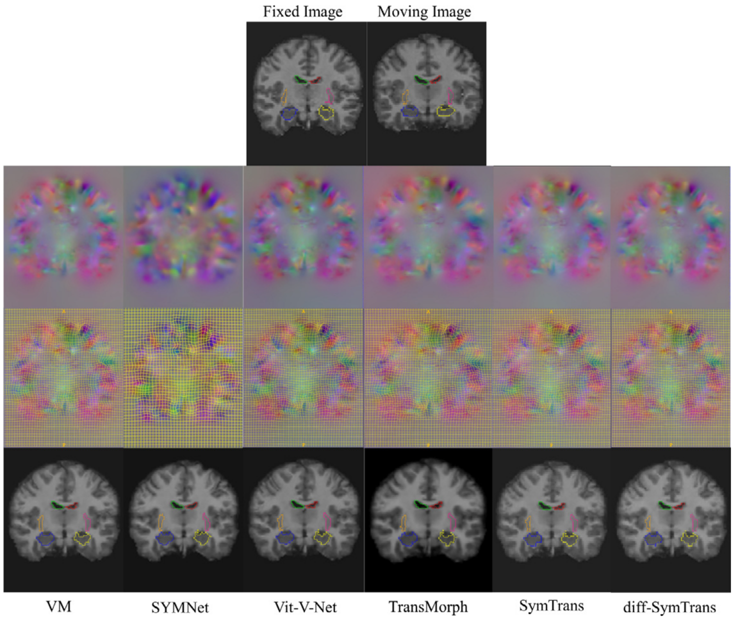

Fig. 4. The atlas-based registration of lateral-ventricle, thalamus, and hippocampus by the VoxelMorph, SYMNet, Vit-V-Net, TransMorph, and the proposed SymTrans and diff-SymTrans.

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 a Lie group, and can be written as $\phi^{(1)} = \exp(v)$. Starting from the initial deformation field at time 0, i.e., the output velocity field from SymTrans, can be formulated as

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

where p is a map of spatial locations. The recurrence function for obtaining the deformation field at time 1 can be expressed as

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

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

4.3.2. Optimization

The proposed SymTrans is optimized in an unsupervised manner by evaluating the similarity between aligned and fixed images. As shown in Fig. 1, given an image pair (I_m, I_f), SymTrans estimates the deformation field ϕ. Then, the STN warps I_f to obtain a warped image $\hat{I}_m$ (denoted as $\hat{I}_m = I_m \circ \phi$). We apply the L_2 loss to both 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 [31], from which 270 scans are selected for our experiment. We first resample each scan to 256 × 256 × 256 with an isotropic voxel size of 1 mm × 1 mm × 1 mm. Then, we conduct the standard preprocessing operation to normalize, perform affine transformation, and strip the skull using FreeSurfer [32]. The segmentation maps of each scan, which are regarded as ground-truth information for evaluation, are also obtained through FreeSurfer. Each scan is cropped to 160 × 192 × 224 and then resampled to 96 × 112 × 96. The dataset is split into 200, 34, and 3×36 scans for the training, validation, and test sets, respectively. The validation and test sets are randomly selected from the OASIS dataset. We sequentially, without repetition, combine pairs of scans in the training set to obtain 39,800 permutations of image pairs. These scan pairs are used to train our proposed and the baseline approaches. We conduct the basis atlas-based registration on the test set. Three groups of six and thirty scans are randomly selected as the atlas and moving images for testing the baseline and the proposed methods, respectively, i.e., we test these methods on 3×6×30 = 540 scan pairs.

The baseline methods and proposed methods are evaluated using the Dice similarity coefficients (DSCs), which calculate the

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

Method	DSC	$\|J(\phi)\| \leq 0$
Affine	0.550 (0.069)	–
SyN	0.675 (0.040)	35.516 (82.015)
VoxelMorph	0.739 (0.031)	1307.711 (586.046)
SYMNet	0.724 (0.026)	1119.861 (335.261)
Vit-V-Net	0.745 (0.029)	1441.585 (577.358)
TransMorph	0.751 (0.026)	1445.309 (549.482)
SymTrans	**0.757 (0.025)**	1485.659 (587.560)
diff-SymTrans	**0.751 (0.025)**	**1.064 (5.840)**

overlap between the ground truth segmentation maps and the warped moving image corresponding segmentation maps. We use the negative Jacobian determinant $|J(\phi)| \leq 0$ to represent the folding number. $|J(\phi)| \leq 0$ indicates where the voxels lose topology preservation and violate the property of diffeomorphism when transformed via the deformation field.

5.2. Baseline methods

We compare the proposed SymTrans with five approaches, including one traditional and four deep learning methods. The symmetric image normalization registration method (SyN) is a traditional iterative method for computing the deformation field [20]. We use the SyN implementation in the ANTs [33] toolbox and set the number of iterations to [100,100,100]. The deep learning baseline methods include the CNN-based VoxelMorph [28], the CNN-based SYMNet [24], the transformer-based ViT-V-Net [13] and the Swin-transformer-based TransMorph [11]. We use the publicly available implementations of these four deep learning methods. We train VoxelMorph, SYMNet, Vit-V-Net and TransMorph with the suggested hyperparameter settings with the same dataset split.

5.3. Implementation details

The proposed framework is implemented using PyTorch. The STN in our method is the same as that utilized in VoxelMorph, Vit-V-Net, and TransMorph. We set the regularization 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 number of training iterations for the deep learning approaches is 300k.

The detailed configuration of the proposed CEMSA-based transformer during training is as follows: $s = \{24, 16, 12\}$ at the 1/4, 1/8, 1/16 resolution stages; the number of heads is $\{2, 4, 8\}$ at each resolution stage; the patch size is $\{3, 3, 3\}$; and the number of grouped convolution 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 indicate the deformation of each anatomical structure. We quantitatively evaluate the accuracies of the baseline methods and the proposed SymTrans using the DSC metric. The nonpositive Jacobian determinants are utilized to assess the folding number. 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 compared to the baseline methods. Diffeomorphic registration using SymTrans (denoted diff-SymTrans) still yields a higher average DSC than the baseline methods and substantially decreases the average folding number, which guarantees the topology of the original moving image. In addition, the lower standard deviations of SymTrans and diff-SymTrans show the strong stability of the proposed SymTrans.

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.

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	18.69	63.61

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 as follows: brainstem (BS), thalamus (Th), cerebellum cortex (CblmC), lateral ventricle (LV), cerebral white matter (CeblWM), cerebellum white matter (CblmWM), putamen (Pu), pallidum (Pa), ventral DC (VDC), caudate (Ca), 4th ventricle (4V), 3rd ventricle (3V), amygdala (Am), hippocampus (Hi), cerebral cortex (CeblC), accumbens (Ac), choroid plexus (CP), inf-lateral ventricle (ILV), and vessel (Ve). The proposed SymTrans outperforms the compared registration approaches on all 19 combined structures. The diff-SymTrans method yields better results than all baseline methods except TransMorph and SymTrans while producing minimal folding. In summary, the proposed symmetric transformer model based on the CEMSA achieves the best results.

5.4.2. Computational complexity

To evaluate the effectiveness of the proposed CEMSA, we compare it with the baseline approaches in terms of numbers of parameters and FLOPs. Table 2 shows the number of FLOPs and parameters of each method. The CNN-based networks, namely, VoxelMorph and SYMNet, have fewer parameters and FLOPs than the transformer-based models because the transformer-based models have many linear operations, which increases the numbers of parameters and FLOPs. Among the three transformer-based methods, our SymTrans achieves the lowest number of FLOPs. Compared with Vit-V-Net and TransMorph, the parameters of SymTrans are much fewer, and there are fewer FLOPs. Specifically, Vit-V-Net employs 12 transformer blocks at the bottom of the model, and each block contains 1.76M parameters. TransMorph employs Swin-Transformer blocks at the 1/4 resolution stage, which model the input embedding patches with 96 dimensions. In the SymTrans encoder, the depth of the CEMSA-based transformer at each resolution stage is equal to the depth of the Swin transformer in TransMorph.

In general, the number of parameters increases with increasing size of the input token. SymTrans applies CEMSA-based transformer blocks at the 1/4, 1/8, and 1/16 resolution levels in the

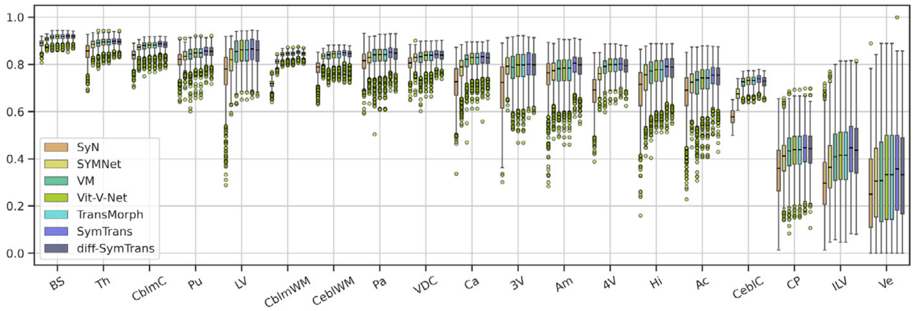

Fig. 5. A boxplot illustrating the DSC values of SyN, VoxelMorph, SYMNet, Vit-V-Net, and ours on each anatomical structure segmentation. We averaged the DSC values of the left and right brain hemispheres and combined them into one structure for visualization.

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

Method	Modules	Encoder	Bottom	Decoder	DSC
E-SymTrans	Patch Emb	✓	✗	✗	0.747 (0.026)
	Fusion	–	–	✗	
	Patch Exp	–	–	✗	
	CETB	✓	✗	✗	
B-SymTrans	Patch Emb	✗	✓	✗	0.735 (0.029)
	Fusion	–	–	✗	
	Patch Exp	–	–	✗	
	CETB	✗	✓	✗	
D-SymTrans	Patch Emb	✗	✗	✓	0.743 (0.028)
	Fusion	–	–	✓	
	Patch Exp	–	–	✓	
	CETB	✗	✗	✓	
SymTrans	Patch Emb	✓	✓	✓	0.753 (0.025)
	Fusion	–	–	✓	
	Patch Exp	–	–	✓	
	CETB	✓	✓	✓	

framework. Even with the application of CEMSA-based transformer blocks at many resolution stages, SymTrans has approximately 49% fewer parameters than Vit-V-Net and 67% fewer parameters than TransMorph. In practice, the GPU memory occupied during training is approximately 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 approximately 6 GB and 7 GB, respectively, of GPU memory with an input padded image size of 96 × 128 × 96. Statistical results regarding the numbers of parameters and FLOPs indicate that the proposed CEMSA is a feasible approach for reducing the number of parameters, which provides a basis for applying 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 ablations are utilized to perform displacement field registration. We train the ablation variants for 100k iterations. Then, we find the best weights on the validation set and test the corresponding variants on the test set.

Table 3 reports the DSC results of three variants of SymTrans. To clearly illustrate the differences among these variant structures, the detailed structural components of each method are listed in Table 3. E-SymTrans contains CEMSA-based transformer blocks in the encoder and replaces CEMSA-based transformer blocks with convolutional blocks in the decoder. In D-SymTrans, only CEMSA-based transformer blocks are utilized in the decoder, and the remaining blocks, 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 a CNN-based architecture that applies 10 CEMSA-based transformer blocks at the bottom. Each convolutional block is followed by a LeakyReLU activation to construct a Conv block. The depths of the Conv blocks are the same as the depths of the replaced CEMSA-based transformer blocks. *Patch expanding* blocks are replaced with the deconvolution operation. The structural forms of E-SymTrans and D-SymTrans correspond to the structural forms of TransMorph and Vit-V-Net. We observe that the original SymTrans achieves the best performance. The results of these ablation variants show that employing the CEMSA-based transformer at the high-resolution levels of the network and applying it symmetrically as an encoder and decoder enhance the registration accuracy. This demonstrates that modeling high-resolution feature maps with a symmetric architecture can help the model recognize meaningful semantic correspondences to anatomical structures.

6. Conclusions

This paper proposed a CEMSA mechanism for capturing local spatial context and reducing semantic ambiguity and parameter quantity. Based on the proposed CEMSA, we built SymTrans for deformable image registration, which takes advantage of the long-range spatial relevance for feature enhancement. The transformer blocks based on CEMSA are applied not only at the bottom but also at the higher-resolution levels in both the encoder and decoder. The qualitative and quantitative evaluation results demonstrate that SymTrans promotes the semantically meaningful correspondence of anatomical structures and provides state-of-the-art registration performance. Furthermore, the ablation studies demonstrate the performance impact of applying the transformer to different components (i.e., the encoder and decoder) of the model, which indicates the effectiveness of the symmetric scheme and the importance of building transformers at the high resolution levels.

CRediT authorship contribution statement

Mingrui Ma: Methodology, Writing – original draft, Validation. **Yuanbo Xu:** Writing – review & editing. **Lei Song:** Visualization. **Guixia Liu:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Nature Science Foundation of China [grant number 61772226]; Science and Technology Development Program of Jilin Province, China [grant number 20210204133YY]; The Natural Science Foundation of Jilin Province, China (Grant number No. 20200201159JC); Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China, Jilin University.

References

[1] T. Gerig, K. Shahim, M. Reyes, T. Vetter, M. Lüthi, Spatially varying registration using gaussian processes, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2014, pp. 413–420.

[2] Y. Wu, W. Ma, M. Gong, L. Su, L. Jiao, A novel point-matching algorithm based on fast sample consensus for image registration, IEEE Geosci. Remote Sens. Lett. 12 (1) (2014) 43–47.

[3] X. Lin, Q. Zou, X. 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, 2021, pp. 1104–1110.

[4] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, C.-W. Lin, Coarse-to-fine CNN for image super-resolution, IEEE Trans. Multimed. 23 (2021) 1489–1502.

[5] G. Zhang, Q. Ma, L. Jiao, F. Liu, Q. 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, 2021, pp. 789–796.

[6] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.

[7] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[10] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[11] J. Chen, Y. Du, Y. He, W.P. Segars, Y. Li, E.C. Frey, Transmorph: Transformer for unsupervised medical image registration, 2021, arXiv preprint arXiv:2111.10480.

[12] Y. Zhang, Y. Pei, H. Zha, Learning dual transformer network for diffeomorphic registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 129–138.

[13] J. Chen, Y. He, E.C. Frey, Y. Li, Y. Du, ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration, 2021, arXiv:2104.06468.

[14] J. Chen, Y. Du, Y. He, W.P. Segars, Y. Li, E.C. Frey, Transmorph: Transformer for unsupervised medical image registration, 2021, arXiv preprint arXiv:2111.10480.

[15] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, X. Xie, After-Unet: Axial fusion transformer unet for medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3971–3981.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[17] Q. Zhang, Y.-B. Yang, Rest: An efficient transformer for visual recognition, Adv. Neural Inf. Process. Syst. 34 (2021).

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

[19] D. Shen, C. Davatzikos, HAMMER: Hierarchical attribute matching mechanism for elastic registration, IEEE Trans. Med. Imaging 21 (11) (2002) 1421–1439, http://dx.doi.org/10.1109/TMI.2002.803111.

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

[21] M.F. Beg, M.I. Miller, A. Trouvé, L. Younes, Computing large deformation metric mappings via geodesic flows of diffeomorphisms, Int. J. Comput. Vis. 61 (2) (2005) 139–157.

[22] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, X. Pennec, SVF-Net: Learning deformable image registration using shape matching, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 266–274.

[23] J. Fan, X. Cao, Z. Xue, P.-T. Yap, D. Shen, Adversarial similarity network for evaluating image alignment in deep learning based registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 739–746.

[24] T. Mok, A. Chung, Fast symmetric diffeomorphic image registration with convolutional neural networks, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020.

[25] A.V. Dalca, G. Balakrishnan, J. Guttag, M.R. Sabuncu, Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces, Med. Image Anal. 57 (2019) 226–236.

[26] J. Wang, M. 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, 2020.

[27] R. Liu, Z. Li, Y. Zhang, X. Fan, Z. 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, 2021, pp. 723–730.

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

[29] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, CVT: Introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22–31.

[30] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, Adv. Neural Inf. Process. Syst. 28 (2015) 2017–2025.

[31] D.S. Marcus, T.H. Wang, J. Parker, J.G. Csernansky, J.C. Morris, R.L. Buckner, Open access series of imaging studies (OASIS): Cross-sectional MRI data in Young, middle aged, nondemented, and demented older adults, J. Cogn. Neurosci. 19 (9) (2007) 1498–1507, arXiv:https://direct.mit.edu/jocn/article-pdf/19/9/1498/1936514/jocn.2007.19.9.1498.pdf.

[32] B. Fischl, Freesurfer, NeuroImage 62 (2) (2012) 774–781, http://dx.doi.org/10.1016/j.neuroimage.2012.01.021, URL https://www.sciencedirect.com/science/article/pii/S1053811912000389, 20 YEARS OF fMRI.

[33] B.B. Avants, N.J. Tustison, G. Song, P.A. Cook, A. Klein, J.C. Gee, A reproducible evaluation of ANTs similarity metric performance in brain image registration, Neuroimage 54 (3) (2011) 2033–2044.