Dual attention network for unsupervised medical image registration based on VoxelMorph

Yong-xin Li, Hui Tang, Wei Wang, Xiu-feng Zhang & Hang Qu

An accurate medical image registration is crucial in a variety of neuroscience and clinical studies. In this paper, we proposed a new unsupervised learning network, DAVoxelMorph to improve the accuracy of 3D deformable medical image registration. Based on the VoxelMorph model, our network presented two modifications, one is adding a dual attention architecture, specifically, we model semantic correlation on spatial and coordinate dimensions respectively, and the location attention module selectively aggregates the features of each location by weighting the features of all locations. The coordinate attention module further puts the location information into the channel attention. The other is introducing the bending penalty as regularization in the loss function to penalize the bending in the deformation field. Experimental results show that DAVoxelMorph achieved better registration performance including average Dice scores (0.714) and percentage of locations with non-positive Jacobian (0.345) compare with VoxelMorph (0.703, 0.355), CycleMorph (0.705, 0.133), ANTs SyN (0.707, 0.137) and NiftyReg (0.694, 0.549). Our model increases both model sensitivity and registration accuracy.
CNN and spatial transformation (ST) exceeded others\textsuperscript{12}. It showed higher computational efficiency since it does not require a large amount of labeling data. It is a probabilistic generative model that defines registration as a parametric function and models functions using CNN and reconstruct images using ST layer. It is an inference algorithm based on unsupervised learning to provide diffeomorphic guarantees and uncertainty estimates, which learned the parameters by optimizing the variational lower bound\textsuperscript{12–15}. However, we note that VoxelMorph fail to effectively suppress useless information on the spatial field during modeling the registration parameter function. Hence, in order to improve sensitivity issue of VoxelMorph, we propose a modified unsupervised learning model by employing the attention mechanism in the registration field to focus on important features and suppress unnecessary ones.

Our novel unsupervised learning model DA-VoxelMorph for deformable medical image registration have two contributions. First, our CNN architecture module that combines the dual attention mechanism of coordinate attention and spatial attention. The module is superior in automatically learning different shapes and sizes of the target structure, implicitly learning to suppress irrelevant regions in the image during model training, and highlighting significant useful features for the registration task. Secondly, we introduce a modified loss function, which consists of cross-correlation and bending penalty regularization method. Our loss function has superior robustness and less susceptible to the effects of image grayscale distribution and contrast.

**Related work**

**Deep learning-based registration.** Depending on the type of annotation available in the training data, deep learning registration approaches can be broadly categorized as supervised, weakly supervised, and unsupervised transformation estimation\textsuperscript{16}. The supervised learning requires the training datasets to include ground-truth deformation field, which is obtained either by simulating deformations and deformed images, or by running classical registration methods on pairs of scans. Based on a patch-based training system, Yang et al. design a deep encoder–decoder network to initialize the momentum of the large deformation diffeomorphic metric mapping registration model\textsuperscript{17}. Sokooti et al. trained a 3D CNN to register chest CT data using artificially generated displacement vector field\textsuperscript{18}. On the other hand, image-to-image prediction can be performed by a fully convolutional neural network (FCN), in which pixel-level features are predicted. For example, Fischer et al. proposed a novel CNN model for optical flow prediction, which trained end-to-end on a synthetic dataset and performed image-to-image optical flow prediction\textsuperscript{19}.

Alternatively, some works have focused on the weakly supervised learning. For example, Hu et al.\textsuperscript{20}, Xu and Niethammer\textsuperscript{21} showed networks trained to maximize the alignment between tissue labels. Besides, Blendowski et al. used a shape encoder–decoder network to extract cardiac shape representations as a basis for registration\textsuperscript{22}. Drawbacks included the time-consuming nature of tissue labeling and the dependence on the performance of the resulting network on the accuracy of tissue labeling are well recognized\textsuperscript{23}.

Although supervised methods have presented a promising direction, ground truth warp fields derived by traditional registration tools as ground truth can be laborious to acquire and can restrict the learned deformation types. In contrast, unsupervised learning mainly uses spatial transformer networks (STN) to warp moving image with estimated registration field, and training of the estimators relies on the design of data similarity function and smoothness of estimated registration field\textsuperscript{24}. First unsupervised learning methods include Deep Learning Image Registration (DLIR) proposed by de Vos et al.\textsuperscript{2} and non-rigid image registration using FCN introduced by Li and Fan\textsuperscript{25}. In addition, the starting point of the present work is the VoxelMorph framework. It is considered as the state-of-the-art, as it is fully unsupervised and allows for a clinically feasible real-time solution by registering full 3D volumes in a single shot\textsuperscript{12}. Given the large dataset the VoxelMorph model has been evaluated, it exceeded other medical registration methods. With an appropriate loss function such as mutual information, the model can perform multimodal registration.

**Attention mechanism.** Attention mechanisms tells a model “what” and “where” to attend and have been proven helpful in a variety of computer vision tasks\textsuperscript{26}, such as image classification\textsuperscript{27,28}, and image segmentation\textsuperscript{29,30}. Wang et al. introduced an encoder–decoder style attention module\textsuperscript{31}. This high-capacity unit is inserted into deep residual networks between intermediate stages. In contrast, Hu et al. proposed the SE block, which is a lightweight gating mechanism. It specialized to model channel-wise relationships in a computationally efficient manner and designed to enhance the representational power of basic modules throughout the network\textsuperscript{28}. However, the SE attention neglects the importance of positional information, which is critical to capturing object structures in vision tasks\textsuperscript{32}. To exploit positional information, later works included BAM and CBAM attempt to reduce the channel dimension of the input tensor and then compute spatial attention using convolutions\textsuperscript{33,34}. Given that convolutions can only capture local relations but fail to model long-range dependencies that are essential for vision tasks, Hou et al. proposed an efficient attention mechanism coordinate attention by embedding positional information into channel attention to enable mobile networks to attend to large regions while avoiding incurring significant computation overhead\textsuperscript{35}. The coordinate attention block is another starting point for the present work. In our AttentionVoxelMorph network, we introduce Dual Attention CNN Architecture by combining coordinate attention block and spatial attention block to further strengthen salient features and suppress useless information in the registration field.

**Loss function of image registration model.** The loss function of a non-rigid image registration model based on unsupervised learning is usually composed of two parts. One part is the similarity measure between the reference image and the deformed floating image. The other is the spatial regularization of the deformation field predicted by the network to constrain the spatial smoothness of the deformation field. There are three commonly used image similarity measures: mean squared voxel difference, cross-correlation and mutual informa-
tion. Mean Squared Voxel Difference and cross-correlation are usually used for unimodal images. The mutual information is usually used for multimodal images, which has better robustness in unimodal images. In the processing of network training, discontinuous deformation fields are often generated in the network in order to measure the similarity of images to the maximum extent. Therefore, it is usually necessary to apply spatial smoothing constraints on the predicted deformation field, that is, to penalize the spatial gradient of the deformation field, such as spatial regularization in VoxelMorph to calculate the square of the L2 norm of the gradient of the deformation field. Recent work has proposed a regularization method called bending penalty, which computes the second-order gradient of the deformation field to penalize folding in the deformation field, and we will incorporate this into our loss function.

Methods
On the basis of VoxelMorph framework, we propose an VoxelMorph Dual Attention CNN Architecture, an attention enhanced approach that further inhibit the useless information in the spatial field and improve the model accuracy. We learn the network parameters in an unsupervised fashion. We combine the attention modules that generate inter-spatial relationship, consider both positional information and channel-wise relationships. We confirm that all methods were carried out in accordance with relevant guidelines and this study was approved by the Ethics Committee of the Affiliated Hospital of Yangzhou University (2017-YKL11-021).

VoxelMorph CNN architecture. VoxelMorph is an unsupervised registration framework based on convolutional neural network (CNN). It estimates the dense deformation field in one step by cascading U-Net and STN structures to realize deformable registration of 3D brain MRI images. Under unsupervised conditions, the registration accuracy can be significantly better than SyN, which confirmed the superiority of unsupervised registration method over supervised and traditional registration methods. We use the same VoxelMorph CNN architecture proposed by Balakrishnan et al. The parameterization of $g_\theta(f, m)$ is based on a CNN convolutional neural network architecture that consists of encoder and decoder sections with skip connections. The VoxelMorph CNN architecture concatenates the moving image $M$ and fixed image $F$ into a 2-channel 3D image as the input and generate the corresponding $\phi$ (Fig. 1). In the encoder stage, $3 \times 3 \times 3$ convolutions with stride 2 followed by Leaky ReLU activations are used to reduce the spatial dimensions until the bottleneck layer. In the decoder stage, we alternate between upsampling, convolutions and concatenating skip connections that propagate features learned during the encoding stages directly to layers generating the registration.

Dual attention CNN architecture. In the medical images, attention needs to be focused on salient features (relevant tissues or organs) that are useful for a specific task, suppressing irrelevant regions in the input image. In cascaded neural networks, an explicit external tissue/organ localization module is required, and the use of spatial attention to learn individual elements with respect to the target can replace this operation. In order to make the most use of the spatial information extracted from the encoding and corresponding decoding stages, we propose a CNN architecture module that integrates the dual attention mechanism of coordinate attention and spatial attention.

Coordinate attention. Note that the standard convolution itself is difficult to model the channel relationships. However, to aggregate global information, global average pooling has been commonly adopted. In order to get channel-wise statistics, we first apply Squeeze- and- Excitation (SE) blocks proposed by28. Given the input X, the squeeze step for the $c$-th channel is calculated as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j)$$

where $z_c$ is the output related to the $c$-th channel. The input comes directly from a convolutional layer with a fixed kernel size and then be considered as a collection of local descriptors. Noticeably, the global pooling operation squeezes global spatial information into channel descriptors, causing the difficulty in preserving positional...
Therefore, we propose to introduce the coordinate attention blocks, demonstrating in Fig. 2. It is demonstrated by Hou et al.\textsuperscript{35}, which considers both inter-channel relationships and positional information.

Two 1D global pooling operations are used to aggregate the input features into two separate directional perceptual feature maps along the vertical and horizontal directions respectively. Each attention map captures the long-distance dependence of the input feature map along one spatial direction. As shown in Fig. 2, the given input consists of two spatial extents of pooling kernels (H, 1) or (1, W) to encode each channel along the horizontal coordinate (X) and the vertical coordinate (Y), respectively. The output of c-th channel at height h can be formulated as Eq. (2), and the output of the vertical coordinate can be formulated as Eq. (3):

\[
z^h_c(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i)
\]  

\[
z^w_c(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w)
\]

**Figure 2.** The coordinate attention block, where X Avg Pool and Y Avg Pool represent 1D horizontal global pooling and vertical global pooling, respectively.
The two transformations described above aggregate characteristics along the two spatial directions, resulting in a pair of direction-aware feature maps. Demonstrated in the Fig. 2, X Aveg Pooling and Y Aveg Pooling represents that the two transformations that enabling our attention block to capture long-range dependencies along one spatial dimension while preserving exact positional information along the other, which allow networks to detect objects of interest accurately.

After the coordinate information embedding stage mentioned above, we perform the coordinate attention generation by concatenating the aforementioned two transformations first and then use 1 × 1 convolutional transformation function $F_1$:

$$f = \delta(F_1(\left[z^h, z^w\right]))$$  \hfill (4)

where $[\cdot, \cdot]$ is the operation of concatenating along the spatial dimension, $\delta$ is the non-linear activation function, and $f \in \mathbb{R}^{C/r \times (H \times W)}$ is the intermediate feature map that encodes spatial information in both the horizontal and the vertical directions. $r$ is the reduction ratio for controlling SE block. Then, splitting $f$ along the spatial dimension into two separate tensors: $f^h \in \mathbb{R}^{C/r \times H}$ and $f^w \in \mathbb{R}^{C/r \times W}$. Apply another two 1 × 1 convolutional transformations $F_h$ and $F_w$ to transform $f^h$ and $f^w$ into tensors with same channel numbers to the input X. Equations (5) and (6) show as below:

$$g^h = \sigma(F_h(f^h))$$  \hfill (5)

$$g^w = \sigma(F_w(f^w))$$  \hfill (6)

where $\sigma$ is the sigmoid activation function. To reduce overhead model complexity, here, we usually reduce number of f channel with appropriate reduction ratio $r$. Then we expand the outputs $g^h$ and $g^w$ as attention weights. Finally, the output of Coordinate Attention Block as the following Eq. (7):

$$y_c(i, j) = x_c(i, j) \times g^h_c(i) \times g^w_c(j)$$  \hfill (7)

**Spatial attention.** The pooling layer in the ordinary convolutional neural network directly uses max pooling or average pooling to compress the image information and reduce the amounts of operations to improve the accuracy rate. We need to further suppress the irrelevant regions in the input image and highlight the significant features of specific local areas, so we introduce the spatial attention module to further extract the key information.

To compute spatial attention, as in CBAM\textsuperscript{34}, we apply the following Attention module schematic (shown as Fig. 3). We first apply average-pooling and max-pooling operations along the channel axis and concatenate them together and apply ReLU function to obtain $\sigma_1(W_x^T x^j + W_g^T g_i + b_g)$, which refers to $F_{int} \times H_g W_g D_g$ in the Fig. 3. Then, we apply another 1 × 1 convolutional computation to get $q_{att}$, the formulation is shown below:

$$q_{att} = \psi^T(\sigma_1(W_x^T x^j + W_g^T g_i + b_g)) + b_\psi$$  \hfill (8)

where $W_x \in \mathbb{R}^{F_x \times F_{int}}$, $W_g \in \mathbb{R}^{F_g \times F_{int}}$, $\psi \in \mathbb{R}^{F_{att} \times 1}$ and bias terms $b_\psi \in \mathbb{R}$, $b_g \in \mathbb{R}^{F_{int}}$ are computed using channel-wise 1 × 1 × 1 convolutions for the input tensors. Finally, using sigmoid activation function on $q_{att}$ to get attention weight $\alpha$, the formula is shown below:
This part is derivable and the value of attention coefficients can be adjusted by training. The formulation is shown below:

\[ \alpha_{ll}^{l} = \sigma_{\text{att}} \left( \tilde{\alpha}_{ll}^{l}, g_{l}; \Theta_{\text{att}} \right) \]  

(9)

This part is derivable and the value of attention coefficients can be adjusted by training. The formulation is shown below:

\[ \frac{\tilde{a}(\tilde{x}_{l}^{l})}{\tilde{a}(\Phi^{l-1})} = \frac{\tilde{a}(\tilde{a} f(\tilde{x}_{l}^{l}; \Phi^{l-1}))}{\tilde{a}(\Phi^{l-1})} = \alpha_{ll}^{l} \frac{\tilde{a}(f(\tilde{x}_{l}^{l}; \Phi^{l-1}))}{\tilde{a}(\Phi^{l-1})} + \frac{\partial (\alpha_{ll}^{l})}{\partial (\Phi^{l-1})} \tilde{x}_{l}^{l} \]  

(10)

Figure 4. The dual attention CNN architecture. Each rectangle represented a 3D volume, generated from the previous volume by a 3D convolutional network layer. The spatial resolution of each volume was printed below. Fusion coordinate attention block and spatial attention block were added to the hop connection between each encoding stage and the corresponding decoding stage.

DAVoxelMorph. By adding the Dual Attention CNN Architecture into the standard VoxelMorph framework, the overview of DAVoxelMorph is demonstrated in the Fig. 5. To be specific, we propose to apply Dual Attention CNN Architecture to build registration field \( \phi \) from the mapping of \( f \) to \( m \), where \( f \), \( m \) are two inputs of image volumes from n-dimensional space, \( \alpha \) denotes the displacement field. We model a function \( g_{\theta}(f, m) = u \) using a convolutional neural network (CNN), where \( \theta \) are network parameters. The registration field \( \phi \) is stored in a \( n + 1 \)-dimensional image. In other words, for each \( p \in \Omega, u(p) \) is a displacement such that \( f(p) \) and \( m \circ \phi(p) \) correspond to similar anatomical locations.

The network of \( \alpha_l \) in the registration field so that voxel in \( m \) and \( f \) can correspond to similar anatomical locations. Similarly, \( f(p) \) and \( m \circ \phi(p) \) denote the anatomic segmentation. Then, the network takes \( f \) and \( m \) as input, and computes \( \phi \) using a set of parameters \( \theta \). We warp \( m \) to \( m \circ \phi \) using a spatial transformation function, enabling evaluation of the similarity of \( m \circ \phi \) and \( f \).

Loss functions. Our overall loss function \( \mathcal{L}_{\text{ul}}(\cdot, \cdot, \cdot) \) uses input volumes and generated registration filed to evaluate the model. It consists of two components, including: (1) \( \mathcal{L}_{\text{sim}} \) to penalize differences of appearances; (2) \( \mathcal{L}_{\text{smooth}} \) to penalize local spatial variations in \( \phi \):

\[ \mathcal{L}_{\text{ul}}(f, m, \phi) = \mathcal{L}_{\text{sim}}(f, m \circ \phi) + \mathcal{L}_{\text{smooth}} \]  

(11)
Let \( \hat{f}(p) \) and \( [\hat{m} \circ \phi](p) \) denote local mean intensity images: \( \hat{f}(p) = \frac{1}{n} \sum_{p_i} f(p_i) \). The local cross-correlation of \( f \) and \( m \circ \phi \) is written as:

\[
CC(f, m \circ \phi) = \frac{\left( \sum_{p_i} f(p_i) - \hat{f}(p) \right) \left( [m \circ \phi](p_i) - [\hat{m} \circ \phi](p) \right)}{\left( \sum_{p_i} \left( f(p_i) - \hat{f}(p) \right)^2 \right) \left( \sum_{p_i} \left( [m \circ \phi](p_i) - [\hat{m} \circ \phi](p) \right)^2 \right)}^2
\]

A higher CC indicates a better alignment, yielding the loss function

\[
L_{\text{sim}}(f, m \circ \phi) = -CC(f, m \circ \phi).
\]

During the training of networks, previous studies used smoothness constraint of its spatial gradient\(^{38,39}\). However, we introduced a bending penalty term which regularizes the transformation,

\[
L_{\text{smooth}} = \alpha P,
\]

where \( P \) is the affine alignment, \( \alpha = 0 \), and \( \alpha \) is empirically determined to be 0.05 for all deformable image registration.

Based on the general form of such a penalty term has been described by Wahba\(^{40}\), our bending penalty takes the following form:

\[
P = \frac{1}{V} \int_0^X \int_0^Y \int_0^Z \left[ \left( \frac{\partial^2 T}{\partial x^2} \right)^2 + \left( \frac{\partial^2 T}{\partial y^2} \right)^2 + \left( \frac{\partial^2 T}{\partial z^2} \right)^2 + 2 \left( \frac{\partial^2 T}{\partial xy} \right)^2 + 2 \left( \frac{\partial^2 T}{\partial xz} \right)^2 + 2 \left( \frac{\partial^2 T}{\partial yz} \right)^2 \right] dxdydz
\]

**Experiments**

**Datasets and preprocessing.** We chose to experiment on publicly available data to demonstrate the performance of our approach. We evaluated our method and other registration methods on the brain MRI dataset LPBA40. LPBA40 datasets included a total of 56 structures labeled in MRI of 40 healthy, normal volunteers. Standard pre-processing steps were performed, including resampling all scans to a 256 × 256 × 256 grid with 1 mm isotropic voxels, affine special normalization, brain extraction using Freesurfer\(^{41}\), and crop the resulting images to 160 × 192 × 224.

**Evaluation metrics.** We will evaluate our method and other registration methods from two perspectives. Registration mass and deformation field. For registration quality, we used an assessment based on volume overlap between organ segments and quantified it using the Dice score\(^{14}\). Comparing the mean and standard deviation of scores across the various registration methods, formulated as follows:

\[
\text{Dice}(A, B) = 2 \cdot \frac{|A \cap B|}{|A| + |B|}
\]

where \( A \) is the reference image and \( B \) is the target image. A Dice score of 1 indicates that the structures are identical, and a score of 0 indicates that there is no overlap.
To evaluate the regularity of the registration field $\varphi$, the Jacobian matrix captures the local properties of $\varphi$ around voxel $p$. We count all non-background voxels for which $|\det(D\varphi^{-1})| \leq 0$, where the deformation is not diffeomorphic\(^\text{(16)}\). Formula is shown below:

$$\mathcal{N} := \sum \delta(\det(D\varphi^{-1}) < 0)$$

**Implementation.** We compare DAVoxelMorph with various registration methods that represent the most advanced registration performance at present. Two non-deep learning based methods include symmetric standardized SyN\(^\text{(37)}\) and NiftyReg\(^\text{(43)}\) in the publicly available advanced standardization tools (ANTs)\(^\text{(44)}\) software package. Two deep learning based methods, including VoxelMorph\(^\text{(12,38)}\) and CycleMorph\(^\text{(45)}\). There are four methods. Syn and NiftyReg are both non learning based methods. They achieve registration by optimizing the energy function of image pairs and have achieved remarkable success in various computational anatomy studies. VoxelMorph uses unsupervised learning, the loss performance reaches a high level in Dice coefficient, and the training time is greatly reduced. It can use auxiliary information and coarse label information to improve network performance. It is a very classic baseline network in the registration field. CycleMorph uses cyclic consistency to provide an implicit regularization to preserve the topology structure, which overcomes the problem that the registration method of deep learning usually cannot guarantee the preservation of topology.

For VoxelMorph implementation, we implemented our approach using PyTorch on a computer equipped with an Nvidia RTX A2000 GPU and an Intel Xeon Silver 4208 CPU. The Adam optimizer with a learning rate of $10^{-4}$, and a default of 50,000 iterations.

In our experiment, we split LPBA40 dataset into 30 training images and 10 testing images. We randomly choose one image from testing images as fixed image, and input 30 training images to the DAVoxelMorph model. We use Adam Optimizer with a learning rate of $4 \times 10^{-4}$, four scales with a default of 50,000 iterations, NCC as the image similarity loss function, and batch size as 1.

**Results and discussion**

**Ablation study.** To demonstrate the performance of each key module in the DAVoxelMorph model, we perform a series of ablation experiments on the open dataset LPBA40\(^\text{46}\). The corresponding results of each module in DAVoxelMorph are demonstrated in the Table 1, and Fig. 6.

Compared to DAVoxelMorph without bending penalty, DAVoxelMorph with bending penalty showed higher results of Dice score on LPBA 40 dataset, indicating the better performance.

Secondly, we compare the effectiveness of combination of Dual Attention CNN Architecture and bending penalty. Results of Dice score demonstrate that the combination model is superior to other models.

Taken together, bending penalty and Dual Attention CNN Architecture are two essential components in our DAVoxelMorph model, which makes the performance superior to other models.

**Comparison with other methods.** The registration results on LPBA40 dataset with different quantitative evaluation measures are shown in Table 1. We compare our method with various methods.

Through experiments, on the open data set LPBA40 of brain MRI, the dice score and Jacobian matrix of the proposed method are higher than other methods. It is indicated that DAVoxelMorph proposed in this paper is best among all five methods in registration quality and three methods in deformation field (NiftyReg method does not generate deformation field). The configuration results and intermediate processes of the five methods in the experiment are shown in Fig. 6. DAVoxelMorph shows the best registration effect. The dice score and Jacobian matrix obtained by the five methods in the experiment are shown in Table 1. The Dice coefficient of this method is 0.714 (0.127), and the percentage of positions with non-positive Jacobian is 0.345, which is improved compared to the other methods.

**Conclusions**

In our proposed method, the spatial attention mechanism calculates the attention map in the spatial dimension and focuses on the parts of the input image that need to be emphasized or suppressed. The coordinate attention mechanism embeds location information into channel attention to enable mobile networks to pay attention.

| Method                              | Dice      | % of $\mathcal{N}$ |
|-------------------------------------|-----------|---------------------|
| NiftyReg                            | 0.694 (0.122) | 0.549               |
| VoxelMorph                          | 0.703 (0.125) | 0.355               |
| ANTs SyN                            | 0.707 (0.123) | 0.137               |
| CycleMorph                          | 0.705 (0.133) | 0.157               |
| DAVoxelMorph (Bending Penalty only) | 0.709 (0.122) | 0.255               |
| DAVoxelMorph (Dual Attention CNN Architecture only) | 0.712 (0.126) | 0.365               |
| DAVoxelMorph (Dual Attention CNN Architecture and Bending Penalty) | 0.714 (0.127) | 0.345               |

Table 1. Presents result on LPBA 40 dataset, including average Dice scores and percentage of locations with non-positive Jacobian. Standard deviations are presented in parentheses.
in a wider range. The dual attention CNN architecture enables our model to focus on the identification and positioning of tissues and organs that are more critical to the registration task. It also reduces or even filters the attention to other information, so to improve the efficiency and accuracy of registration. The bending penalty calculates the two-step degree of the deformation field. It further improves the registration quality by punishing the folding in the deformation field, promoting the affine transformation of the network locally, and enhancing the continuity and the global smoothness of the deformation field.

In conclusion, we propose a spatial attention enhanced, unsupervised learning method DAVoxelMorph for 3D deformable medical image registration. The results showed our model surpasses the basic VoxelMorph model, CycleMorph, as well as the ANTs SyN and NityReg in model sensitivity and registration accuracy with minimal computational overhead. The Dual Attention CNN Architecture in our model can continuously improve the registration performance under different datasets and training sizes while maintaining computational efficiency. Therefore, our proposed DAVoxelMorph model is a general learning model, but not limited to a particular image type or anatomy. It successfully speeds up the medical image analysis and processing pipelines, which can contribute to the clinical settings.

**Data availability**

The datasets generated during and/or analyzed during the current study are available in the [https://www.loni.usc.edu/research/atlas](https://www.loni.usc.edu/research/atlas).

Received: 9 April 2022; Accepted: 15 September 2022
Published online: 28 September 2022

---

**Figure 6.** Registration results of different methods on LPBA40 dataset. The first column is the fixed image and the second column is the moving image. The third column shows the moved image after registration and in the area shown in the green box, DAVoxelMorph shows a better registration effect in detail. The fourth and the fifth column show the deformation field during the registration and label.
References

1. Fan, J., Cao, X., Xue, Z., Yap, P.-T. & Shen, D. Adversarial similarity network for evaluating image alignment in deep learning based registration. Paper Presented at the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018, Cham (2018).

2. Shan, S., Yan, W., Guo, X., Chang, E. I., Fan, Y. & Xu, Y. J. A. Unsupervised end-to-end learning for deformable medical image registration (2017).

3. Fan, J., Cao, X., Yap, P.-T. & Shen, D. BiRNet: Brain image registration using dual-supervised fully convolutional networks. Med. Image Anal. 54, 193–206. https://doi.org/10.1016/j.media.2019.03.006 (2019).

4. Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q. & Shen, D. Deformable image registration based on similarity-steered CNN regression. Paper Presented at the Medical Image Computing and Computer Assisted Intervention—MICCAI 2017, Cham (2017).

5. de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M. & Išgum, I. End-to-end unsupervised deformable image registration. Paper Presented at the Medical Imaging with Deep Learning (MIDL) 2018, Amsterdam (2019).

6. He, X., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J. & Shen, D. Transformers in medical image analysis: A review. J. Classif. 14(8). https://doi.org/10.48550/arXiv.2202.12165 (2022).

7. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. An unsupervised learning model for deformable medical image registration. Paper Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018.

8. Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q. & Shen, D. Deformable image registration based on similarity-steered CNN regression. Paper Presented at the 3rd International Workshop on Multimodal Learning for Clinical Decision Support, Cham (2017).

9. He, X., Hildebrand, D. G. C., Tobin, W. F., Lee, W. C. A. & Jeong, W. K. seEmNet: Serial-section electron microscopy image registration using a spatial transformer network with learned features. Proceedings of the 3rd International Workshop on Multimodal Learning for Clinical Decision Support 249–257. https://doi.org/10.1007/978-3-6588-9.2_1 (Springer, 2017).

10. Dalca, A. V., Balakrishnan, G., Guttag, J. & Sabuncu, M. R. Unsupervised deformable image registration using fully connected generative neural network. Paper Presented at the Medical Imaging with Deep Learning (MIDL), Cham (2018).

11. Li, Y., Wang, Z., Yin, L., Zhu, Z., Qi, G., Liu, Y. X-Net: A dual encoding decoding method in medical image segmentation. Paper Presented at the International Conference on Machine Learning—ICML 2020 (2020).

12. Haskins, G., Kruger, U. & Yan, P. Deep learning in medical image registration: A survey. Neuroimage 158, 378–396. https://doi.org/10.1016/j.neuroimage.2017.07.008 (2017).

13. Sokooti, H., de Vos, B., Berendsen, F. F., Lelieveldt, B. P. F., Išgum, I. & Staring, M. Nonrigid image registration using multi-scale 3D convolutional neural networks. Paper Presented at the Medical Image Computing and Computer Assisted Intervention—MICCAI 2017, Cham (2017).

14. Fischer, P., Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G. & Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. Published as a Conference Paper at International Conference on Learning Representations—ICLR 2020 (2020).

15. Mach. Vis. Appl. 10. 1007/ s00138- 020- 01060-x (2020).

16. de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M. & Išgum, I. End-to-end unsupervised deformable image registration with fully connected generative neural network. Paper Presented at the Medical Imaging with Deep Learning (MIDL), Cham (2018).

17. He, X., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J. & Shen, D. Transformers in medical image analysis: A review. J. Classif. 14(8). https://doi.org/10.48550/arXiv.2202.12165 (2022).

18. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. An unsupervised learning model for deformable medical image registration. Paper Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018.

19. Dalca, A. V., Balakrishnan, G., Guttag, J. & Sabuncu, M. R. Unsupervised learning for fast probabilistic diffeomorphic registration. Paper Presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention (2018).

20. Dice, L. R. Measures of the amount of ecologic association between species. Ecology 26(3), 297–302. https://doi.org/10.2307/193249 (1945).

21. Zhu, Y., Zhou, Z., Liao, G. & Yuan, K. A new unsupervised learning method for 3D deformable medical image registration. Paper Presented at the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (2021).

22. Haskins, G., Kruger, U. & Yan, P. Deep learning in medical image registration: A survey. Mach. Vis. Appl. 31(1), 8. https://doi.org/10.1007/s00138-020-01060-x (2020).

23. Wang, X., Kittik, R., Stynar, M. & Niethammer, M. Quicksilver: Fast predictive image registration—A deep learning approach. Neuroimage 158, 378–396. https://doi.org/10.1016/j.neuroimage.2017.07.008 (2017).

24. Tsotsos, J. K. A Computational Perspective on Visual Attention (MIT Press, 2021).

25. Mach. Vis. Appl. 10. 1007/ s00138- 020- 01060-x (2020).

26. Liu, Y. et al. Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. 49, 1–13. https://doi.org/10.1016/j.media.2018.07.002 (2018).

27. Xu, Z., & Niethammer, M. DeepAtlas: Joint semi-supervised learning of image registration and segmentation. Paper Presented at the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Cham (2019).

28. Bloudowski, M., Boulentia, N. & Heinrich, M. P. Multimodal 3D medical image registration guided by shape encoder-decoder networks. Int. J. Comput. Assist. Radiol. Surg. 15(2), 269–276. https://doi.org/10.1007/s11842-019-01047-0 (2020).

29. Han, R. et al. Deformable MR-CT image registration using an unsupervised, dual-channel network for neurosurgical guidance. Med. Image Anal. 75, 102292. https://doi.org/10.1016/j.media.2021.102292 (2022).

30. Iaderberg, M., Simonyan, K. & Zisserman, A. Spatial transformer networks (2015).

31. Li, H. & Fan, Y. J. Non-rigid image registration using fully convolutional networks with deep self-supervision (2017).

32. Tosots, J. K. A Computational Perspective on Visual Attention (MIT Press, 2021).

33. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. & Lu, H. Dual attention network for scene segmentation. Paper Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019).

34. Hou, Q., Zhang, L., Cheng, M.-M., & Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. Paper Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020).

35. Li, H. & Fan, Y. J. Non-rigid image registration using fully convolutional networks with deep self-supervision (2017).

36. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. Voxelmorph: A learning framework for deformable medical image registration. IEEE Trans. Med. Imaging 38(8), 1788–1800. https://doi.org/10.1109/TMI.2018.2897538 (2019).
39. Du, B., Liao, J., Turkbey, B. & Yan, P. Multi-task learning for registering images with large deformation. IEEE J. Biomed. Health Inform. 25(5), 1624–1633. https://doi.org/10.1109/JBHI.2020.3016699 (2021).
40. Wahba, G. Spline Models for Observational Data (SIAM, 1990).
41. Fischl, B. FreeSurfer. NeuroImage 62(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021 (2012).
42. Ashburner, J. J. N. A fast diffeomorphic image registration algorithm. Neuroimage 38(1), 95–113 (2007).
43. Modat, M. et al. Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. 98, 278–284 (2010).
44. Avants, B. B. et al. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54(3), 2033–2044 (2011).
45. Kim, B. et al. Cyclemorph: Cycle consistent unsupervised deformable image registration. Med. Image Anal. 71, 102036 (2021).
46. Shattuck, D. W. et al. Construction of a 3D probabilistic atlas of human cortical structures. Neuroimage 39(3), 1064–1080. https://doi.org/10.1016/j.neuroimage.2007.09.031 (2008).

Acknowledgements
The authors thank Dr. Dou Weiqiang for his help in completing this study.

Author contributions
L.Y.X. dedicated to the formal analysis, data analysis, data curation and original draft preparation. T.H. worked for writing original draft, visualization the result and the conceptualization in this manuscript. W.W. contributed to the resources of project, data curation and proofreading of results. Z.X.F. contributed to methodology and software preparing and data analysis. Q.H. mainly responsible for methodology and software preparing, review and editing of the manuscript, project administration and supervision of the project.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to H.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022