Applied Intelligence (2023) 53:2936–2950
https://doi.org/10.1007/s10489-022-03659-1

Deformable image registration with attention-guided fusion of multi-scale deformation fields

Zhiquan He[1,2,3] · Yupeng He[2] · Wenming Cao[2]

Accepted: 18 April 2022 / Published online: 14 May 2022

Abstract

Deformable medical image registration plays a crucial role in theoretical research and clinical application. Traditional methods suffer from low registration accuracy and efficiency. Recent deep learning-based methods have made significant progresses, especially those weakly supervised by anatomical segmentations. However, the performance still needs further improvement, especially for images with large deformations. This work proposes a novel deformable image registration method based on an attention-guided fusion of multi-scale deformation fields. Specifically, we adopt a separately trained segmentation network to segment the regions of interest to remove the interference from the uninterested areas. Then, we construct a novel dense registration network to predict the deformation fields of multiple scales and combine them for final registration through an attention-weighted field fusion process. The proposed contour loss and image structural similarity index (SSIM) based loss further enhance the model training through regularization. Compared to the state-of-the-art methods on three benchmark datasets, our method has achieved significant performance improvement in terms of the average Dice similarity score (DSC), Hausdorff distance (HD), Average symmetric surface distance (ASSD), and Jacobian coefficient (JAC). For example, the improvements on the SHEN dataset are 0.014, 5.134, 0.559, and 359.936, respectively.

Keywords Deformable image registration · Attention network · Multi-scale feature extraction · Displacement field fusion · Anatomical segmentation

1 Introduction

Deformable medical image registration builds an optimal anatomical alignment between two images and plays a vital role in helping experts diagnose the disease, follow up the diseases' evolution, and decide the necessary therapies regarding the patient's condition [14]. Co-registering MRI brain images before neuro-morphometry analysis is one example [17]. One of the two images is the source or moving image, which is transformed or distorted by the registration to maximally match the other one, i.e., the target or fixed image.

Traditional image registration usually applies image processing techniques such as key points detection, edge extraction, and region segmentation [24, 44, 45], and maximizes a predefined similarity measure between the transformed moving image and the fixed [2]. Unfortunately, solving such an optimization problem usually yields an unsatisfactory computation efficiency and registration accuracy.

In recent years, deep convolutional neural network (CNN) based methods have made an excellent progress in which the trained deep convolutional network takes the source and target image as input, extracts the features, and predicts a spatial transformation field used to warp the moving image toward the target. Among these methods, the unsupervised are the mainstream as they do not need the ground-truth

✉ Zhiquan He
zhiquan@szu.edu.cn

Yupeng He
2070436116@email.szu.edu.cn

Wenming Cao
wmcao@szu.edu.cn

[1] College of Information Engineering, Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen, People's Republic of China

[2] Guangdong Multimedia Information Service Engineering Technology Research Center, Shenzhen, People's Republic of China

[3] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen, People's Republic of China

transformation field to train the model. Instead, the training process minimizes a loss function consisting of multiple constraints such as the pixel-level image similarity between the fixed and the warped moving [4, 5, 9, 25, 49].

Despite the progress so far, there still exists a large room to improve, especially when the deformation or difference between the input two images is large and complex. The first and last column in Fig. 1 shows such an example. Generative adversarial network (GAN), along with extra anatomical segmentation constraint, was proposed to alleviate the problem [25, 26]. However, GAN-based network generally suffers from unstable training. Iterative registration is another attempt that recursively and progressively warps the moving image toward the fixed using a small number of networks cascaded such as VTN [47] and RCINet [49]. However, it is difficult to train such a recursively structured network and control the number of the cascades as increasing the number of cascades does not guarantee the improvement of registration accuracy. Another issue of this method is the computation efficiency.

Unlike the progressive registration, the large image difference motivates us to combine the deformation fields resulting from the image features of different resolutions or scales. On the other hand, we notice the harmful interference from the uninterested regions on the registration accuracy. Most existing methods use the entire image to calculate the image-related constraints without differentiating the importance of different areas. For example, in Fig. 1, the regions of interest are the white areas in the images, while the rest provides little value for the registration.

Based on these observations, we propose a novel attention-guided fusion of multi-scale deformation fields for deformable medical image registration. Specifically, instead of using the dense registration model to internally learn the attention for the regions of interest in the input image, we propose to use a separate deep CNN to predict the attention mask, which is then multiplied with the input to remove the interference areas for subsequent registration. The dense registration network adopts a U-Net [30] structure and produces deformation fields of multiple resolutions. A Deformation Field Spatial Attention (DFSA) module successively combines the fields of lower resolutions with those of higher resolutions using learned attention weights to form the final deformation field. We enhance the attention prediction and the reconstruction accuracy of the anatomical encoder-decoder by designing a new contour loss. Moreover, to improve the texture and structural similarity after registration, we propose to incorporate the image structure similarity index (SSIM) [40] based constraint into the loss function to better guide the model learning.

We have conducted the experiments on three benchmark datasets, and the results have shown a significant improvement over the state-of-the-art in terms of the average Dice similarity score (DSC) [12], Hausdorff distance (HD), Average symmetric surface distance (ASSD), and Jacobian folding coefficient (JAC) [35]. For example, the improvements over GRNet on the SHEN dataset are 0.014, 5.134, 0.559, and 359.936, respectively.

In summary, we have made the following major contributions in this work.

1. We propose a novel attention-guided deformable registration method based on multi-scale deformation fields fusion to improve the registration accuracy, especially for images with large deformations. Specifically, the predicted attention mask removes the interference from

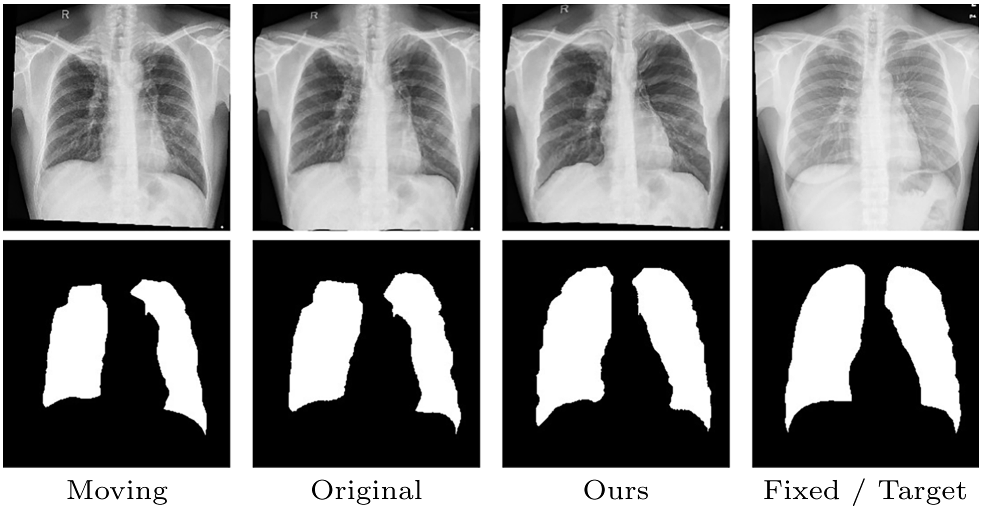


Fig. 1 The first and last columns are the moving and fixed images and the corresponding binary labels. Registration is to find a deformation field to transform the moving image toward the fixed. The second and third columns are the warped moving images from method [26] and ours

the uninterested regions of the input images, and the predicted multiple deformation fields of different scales are combined using the learned attention weight map.

2. We enhance the reconstruction accuracy of the anatomical encoder-decoder by designing a novel contour loss.
3. To improve the image structure similarity after registration, we propose to use the image structural similarity index (SSIM) as a loss term to regularize the model training.
4. We have conducted the experiments on three benchmark datasets, and the results have shown the improvement of registration performance over the state-of-the-art.

The rest of the paper is organized as follows. Section 2 reviews the related works; Section 3 explains the proposed registration method and network structure; Section 4 carries out the comparative experiments to demonstrate the effectiveness of the method, Section 5 discusses the method and results, and Section 6 concludes the paper.

2 Related work

Deformable image registration is usually formulated as an optimization problem that uses pixel displacement fields to represent the spatial transformation and quantifies the similarity between the warped moving image and the fixed. Specifically, the optimization can be defined as

$$\hat{\phi} = \arg\min_{\phi} L(\phi(I_m), I_f) + R(\phi), \tag{1}$$

where I_m, I_f are the moving (source) and fixed (target) image, ϕ is the deformation (or displacement) field, which spatially maps each pixel of I_m to I_f, L is a metric quantifying the alignment quality between the warped moving $\phi(I_m)$ and I_f, and R is a regularization term that imposes some constraints on the transformation field. The optimal transformation $\hat{\phi}$ is obtained through the minimization of (1).

2.1 Image registration based on traditional methods

Traditional deformation registration methods usually extract image features such as key points, edges and region segmentations [14, 24, 44, 45] and optimize the predefined object functions. SimpleElastix by Marstal et al. [27] uses B-spline transformation to parameterize the deformation field and minimizes the image difference iteratively. Similarly, automatic image registration tools SyN [2], ANTs [3], and FAIR [28] define metrics for transformation space and alignment quality and iteratively update the parameters to get the best registration alignment. Intensity-based image features are used to establish optimal registration of the source and target images in vivo imaging experiments for the task of automated detection and tracking of changes in the specimen [22].

One important thread of registration methods adopt diffeomorphic transformation that mathematically is a global one-to-one smooth and continuous mapping with invertible derivatives. Widely used heteromorphic parameterization methods include distance metric mapping [6, 10], DARTEL [1] and diffeomorphic demons [34].

Traditional registration methods mainly use manually curated features to optimize the empirically defined metric, in which the optimization process usually takes a long time to converge, especially when the parameter space to search is high dimensional. Therefore, the registration accuracy is not good enough.

2.2 Image registration based on deep convolutional neural network

With the success of deep convolution neural networks in various computer vision tasks, many deep CNN based methods have been proposed to improve the registration accuracy and efficiency, in which the trained deep convolutional network takes the source and target image as input, extracts the image features and predicts a spatial transformation for registration.

Among these methods, supervised learning methods require the ground truth deformation fields to train the network [16, 29]. The primary issue of these methods is that it is challenging to obtain the high-quality ground truth data. On the contrary, unsupervised deep learning methods are more suitable for practical applications as they train the CNN networks by minimizing the loss function measuring the similarity between the fixed and the warped moving image without the ground truth deformation [23, 31, 35].

In order to improve the registration accuracy, various methods have proposed to incorporate extra information about image modalities and anatomical structures into the registration process. Among these methods, anatomical segmentations have been frequently used. The multi-modal CNN-based image registration method proposed by Hu et al. takes advantage of the provided anatomical labels to infer voxel-level spatial transformation, in which the anatomical segmentations are directly used to calculate the label similarity in the training loss [19]. Similarly, VoxeMorph [5] leverages auxiliary anatomical segmentations during training by adding a Dice coefficient [12] loss term indicating the agreement between the warped moving segmentation and the fixed. Slightly differently, U-ResNet [15] is a multitask network that can generate a deformation field and a segmentation at the same time by sharing and learning the feature representations for both tasks. Global context information about the anatomical segmentations is extracted as one of the loss terms in the training process [26]. Based on this work, Luo et al. combined the anatomical segmentations and GAN framework [25].

It is commonly encountered in medical image registration that the deformation between the input images is large and complex. It would be challenging for a single-step prediction to make accurate registrations for these hard cases. Iterative optimization is a natural option to alleviate this problem. The idea is also widely used in traditional image registration to optimize the objective function [7, 11]. DLIR [35] and VTN [48] designed a stacked network structure with a small number of cascades, where DLIR trained each cascade with the previous stages fixed while VTN jointly trained all the cascades. However, progressive registration was applied in neither training process. Zhao et al.improved the original VTN [48] by considering the iterative registration during the training process [47]. RCINet [49] extended the idea of cascaded network structure to 2D medical image registration and improved the performance using the anatomical segmentations as [26].

2.3 Attention mechanism and image segmentation

Attention mechanism is widely used in deep learning networks for computer vision tasks such as image recognition and semantic segmentation [33, 43], from the spatial attention [36, 42], channel attention [18, 37] to self attention [39, 46]. Spatial attention learns a weight distribution for each spatial location, channel attention assigns weights to feature channels, and a self-attention module computes the response at a position in a sequence or an image by attending to all other positions. Wang et al. proposed a transformer-based [13] method for unsupervised image registration [41] where the deformation fields are learned by the transformer instead of traditional CNNs. Although the transformer-based method has achieved promising performance in several applications, especially natural language processing, it requires a large amount of data and computation to train the model. Instead of internally combining the attention mechanism and the dense registration network, we use the separately predicted segmentation as hard attention to indicate the spatial importance and remove the interference from the uninterested regions. Another benefit of doing this is that we can take advantage of the achievements in image segmentation.

3 Method

In this section, we present the proposed attention-guided fusion of multi-scale deformation fields for deformable image registration.

3.1 Registration framework

As discussed in the previous section, we improve the registration from three aspects, 1) attention prediction for input image indicating its spatial importance and removing the interference from the uninterested regions, 2) combination of deformation fields of multiple scales with learned attention weights, and 3) better regularization of the model training by designing new loss terms. Figure 2 shows the method diagram, which contains two parts.

The upper part is the principle registration network. Let $I_m \in \mathcal{R}^2$ and $I_f \in \mathcal{R}^2$ be the moving and fixed image, respectively. A separate segmentation module predicts the binary mask W_m^a and W_f^a for each image, indicating their regions of interest, e.g., the lung parts. The mask is then multiplied with the input to remove the uninterested areas through element-wise multiplication, namely, $I_m^c = I_m \cdot W_m^a$ and $I_f^c = I_f \cdot W_f^a$. The dense registration network takes I_m^c, and I_f^c concatenated on channels as input to produce the displacement field ϕ, which is the pixel correspondence between the moving and fixed image and used by the spatial transformation network (STN) [20] to generate the warped moving image $\phi(I_m)$. The spatial transformation network is differentiable so that the gradient can be propagated backward during the training process.

The lower part is active only in the training stage, in which the anatomical segmentation mask s_m is warped and used to calculate the constraint terms with respect to the fixed mask s_f to regularize the model learning. The loss function is calculated mainly on the images (L_{img}) and anatomical segmentations (L_{ce} and L_{ae}), as shown in Fig. 2.

3.2 Attention mask prediction

We take the deep CNN model U-Net [30] to predict the attention weight map in Fig. 2. We make two modifications to achieve a better balance between efficiency and accuracy. To speed up the efficiency, we reduce the model parameters from about 7.7 million to 1.9 million. Secondly, by default, the model is trained based on cross-entropy. Our experiments show that the model reports isolated points in the prediction for some cases. To address this issue, we add a new loss term to measure the global-level distance of the predicted binary mask to the ground truth. The loss is defined on the contour of the objects or organs as

$$ct(I_1, I_2) = L_2(\nabla I_1, \nabla I_2), \tag{2}$$

where I_1 and I_2 are two binary mask images, $L_2()$ is the L_2 norm, and ∇ is the Laplacian edge operator [38]. The model must be trained before starting the primary training process

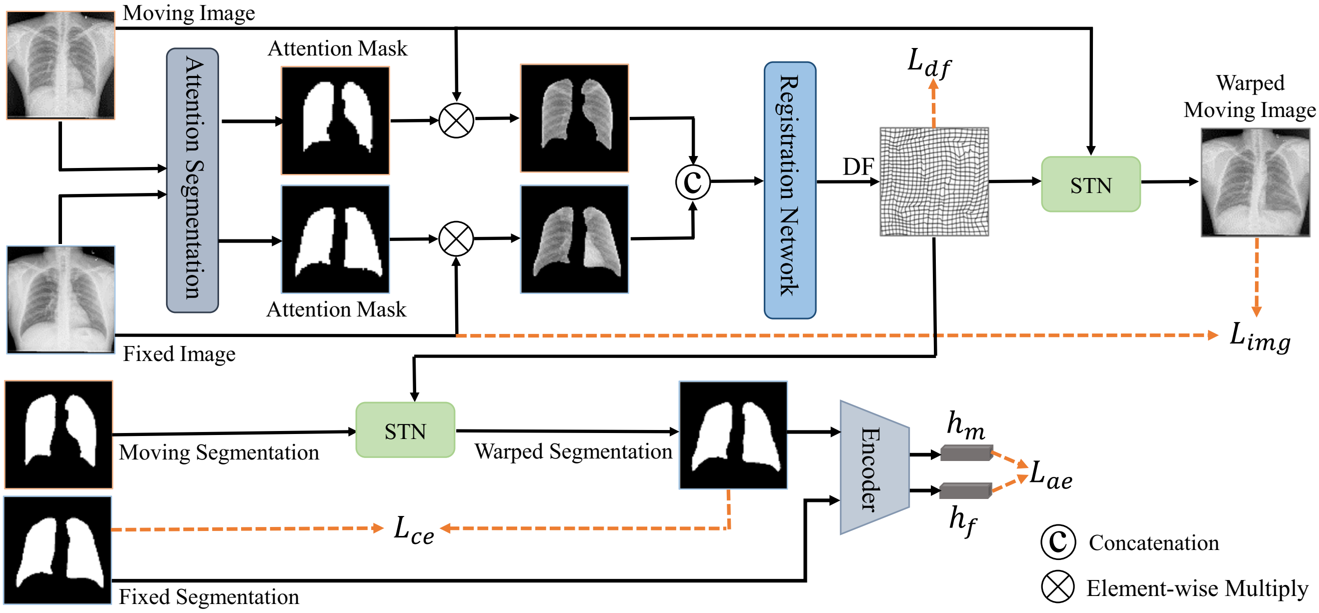


Fig. 2 Overview of the attention-guided fusion of multi-scale deformation fields. In the upper part, a separate segmentation module predicts the binary attention mask for each input to indicate the region of interest, i.e., the lungs in the images. The attention mask is multiplied with the input to remove the uninterested areas. With the cleaned input images, the trained dense registration network produces the deformation field (DF), which the spatial transformer network (STN) uses to warp the moving image. The lower part is active only in the training process, in which the moving anatomical segmentation mask is warped and used to calculate the global context distance with respect to the fixed one. The loss term L_{img} is defined on the warped moving image and the fixed, L_{ce} is defined on the warped moving anatomical segmentation and the corresponding fixed one, L_{ae} is defined on the low-dimensional vectors of the encoder, and L_{df} is defined on the field itself. The dash lines indicate the loss terms

of Fig. 2. The loss function is written as

$$L(I_p, I_t) = ce(I_p, I_t) + \lambda\, ct(I_p, I_t), \tag{3}$$

where I_p and I_t are the predicted and ground truth label images, $ce()$ is the cross-entropy function, and λ is a weight factor.

3.3 Anatomical segmentation encoder

In the lower part of Fig. 2, the warped moving anatomical segmentation and the fixed are input to the encoder to form low-dimensional representation vectors to calculate their distance. The low-dimensional representation is learned through a denoising auto-encoder (DAE), which maps the input image X to a lower-dimensional vector $h = encoder(X)$ by the encoder and then reconstructs X by the decoder $\tilde{X} = decoder(h)$. Training such a model minimizes the reconstruction error of the input [25, 26, 49]. Similar to the attention segmentation prediction, we add a contour loss defined in (2) to improve the global shape of the reconstructed mask. Again, the model should be trained beforehand. And the loss function is

$$L(I, I') = ce(onehot(I), onehot(I')) + \lambda\, ct(I, I'), \tag{4}$$

where I and I' are the input and reconstructed segmentation mask, $onehot()$ is the one-hot coding function, $ce()$ is the cross-entropy function, $ct()$ is the contour loss, and λ is a weight factor.

3.4 Dense registration network

The dense registration network shown in Fig. 3 contains multiple branches corresponding to different feature resolutions or scales. Input images are successively down-sampled with a ratio of 0.5 for branches from top to down to extract features and produce displacement fields of different scales. Let DF_k^c be the deformation field after the last convolution and DF_k^a be the combined deformation field of the k-th branch.

We have

$$DF_k^a = DFSA(DF_{k+1}^a) + DF_k^c, \tag{5}$$

where $k = 1, 2, 3$ is the branch index from top to down and $DFSA()$ is the displacement field spatial attention module. Figure 4 shows how the spatial attention weights are applied to the displacement field. The X and Y channels of the deformation field DF_k^c are processed independently by the convolution layers to obtain weight masks with values ranging in [0, 1]. A new deformation field is formed with the two channels multiplied by the corresponding weight maps. From (5), we have the following channel-wise addition and

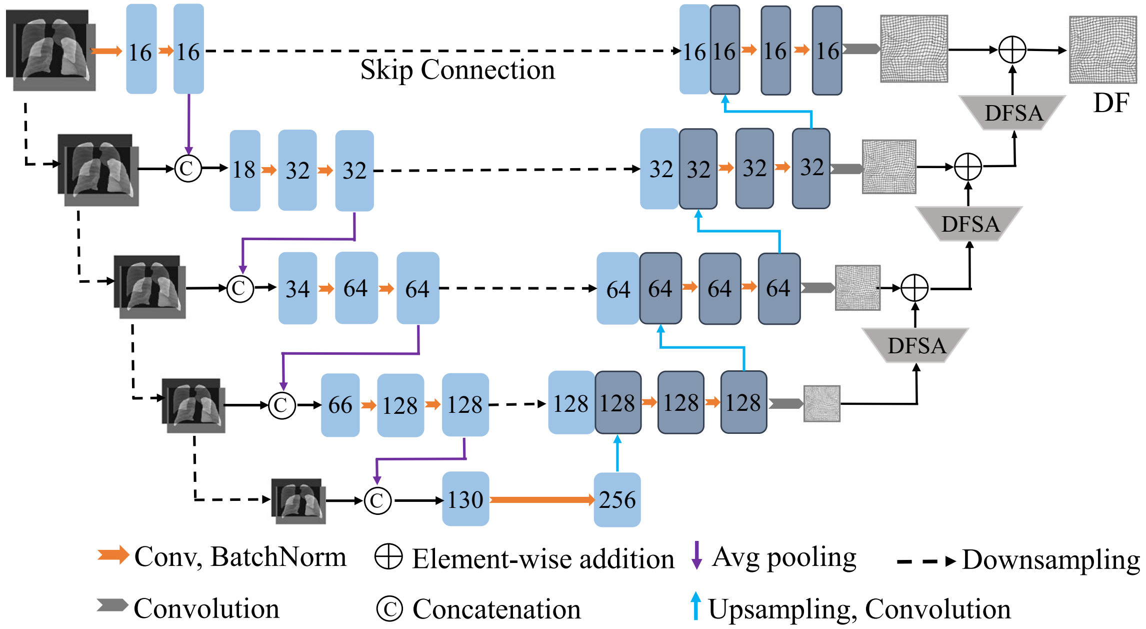


Fig. 3 Dense registration network. Input images are successively down-sampled with a ratio of 0.5 for branches from top to down to extract features and produce displacement fields of different scales. The Displacement Field Spatial Attention block (DFSA) combines the displacement field of one branch with that from the adjacent upper branch. The number of feature channels are shown in the blocks

multiplication

$$DF^a_k.x \;=\; DF^c_{k+1}.x \cdot W_k.x + DF^c_k.x \qquad (6)$$
$$DF^a_k.y \;=\; DF^c_{k+1}.y \cdot W_k.y + DF^c_k.y,$$

where x, y means the X and Y channels.

3.5 Loss function for model training

To train the model in Fig. 2, the entire loss function contains three parts, i.e., the image similarity L_{img}, the anatomical segmentation similarity L_{as}, and the displacement field smoothness L_{df}.

Image Similarity Loss measures the image alignment quality after registration, namely, the similarity between the warped moving image $\phi(I_m)$ and the fixed I_f. To strengthen the image similarity, we propose the following intensity-based and structure-based similarity terms.

$$L_{img} \;=\; \lambda_{ncc}\; NCC(I_f, \phi(I_m)) \qquad (7)$$
$$\qquad +\lambda_{ssim}\; SSIM(I_f, \phi(I_m)),$$

where I_m, I_f are the moving and fixed images, ϕ is the displacement field. The normalized cross-correlation NCC is defined as [4]

$$NCC(I_1, I_2) = \frac{\sum_{x\in\Omega}(I_1(x)-\bar{I}_1)(I_2(x)-\bar{I}_2)}{\sqrt{\sum_{x\in\Omega}(I_1(x)-\bar{I}_1)^2\sum_{x\in\Omega}(I_2(x)-\bar{I}_2)^2}}, \qquad (8)$$

where x is the coordinate index, $\bar{I}_1$ and $\bar{I}_2$ are the mean values. NCC measures the degree of pixel-intensity

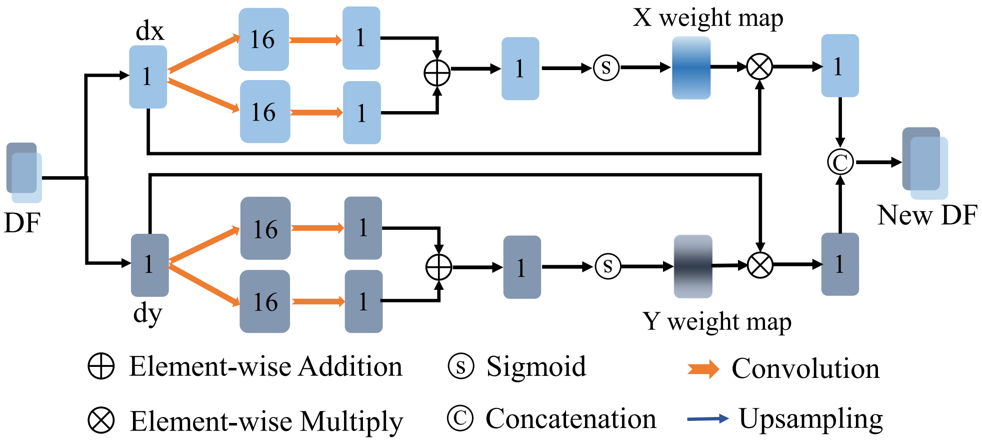


Fig. 4 Deformation field spatial attention (DFSA) network. The X and Y channels of the deformation field are processed independently by the convolution layers to obtain weight masks with values ranging in [0, 1]. A new deformation field is formed with the two channels multiplied by the corresponding weight maps. The numbers in blocks are the number of feature channels

similarity between two images. SSIM calculates the structural similarity index, which is widely used to measure the perceptual quality of images [40]. Let A and B be the two images being compared. A window moves pixel-by-pixel from the top left corner to the bottom right corner of the image. In each step, the local statistics $\Theta(A_j, B_j)$ index is calculated within local window j as follows:

$$\theta(A_j, B_j) = \frac{(2 \cdot m_{A_j} m_{B_j} + C_1) \cdot (2 \cdot \sigma_{A_j B_j} + C_2)}{(m_{A_j}^2 + m_{B_j}^2 + C_1)(\sigma_{A_j}^2 + \sigma_{B_j}^2 + C_2)}, \tag{9}$$

where $m_{A_j}, m_{B_j}, \sigma_{A_j}, \sigma_{B_j}, \sigma_{A_j B_j}$ represent the average intensity of image patches A_j and B_j, the standard deviation of A_j and B_j, and covariance between A_j and B_j, respectively. C_1 and C_2 are two constants of small positive values introduced to avoid numerical instability. The SSIM index between images A and B is defined by

$$SSIM(A, B) = \frac{\sum_{j=1}^{N_s} W(A_j, B_j)\theta(A_j, B_j)}{\sum_{j=1}^{N_s} W(A_j, B_j)}, \tag{10}$$

where N_s is the number of local windows in the image and $W(A_j, B_j)$ is the weights applied to window j [40]. In this work, we use a Pytorch model (https://github.com/aserdega/ssim-pytorch) to approximate the SSIM function.

Anatomical Segmentation Loss regularizes the training process by measuring the distance between the warped moving segmentation and the fixed one.

$$L_{as} = \lambda_{ce} \; L_{ce}(s_f, \phi(s_m)) + \lambda_{ae} \; L_{ae}(s_f, \phi(s_m)), \tag{11}$$

where s_m and s_f are the moving and the fixed anatomical segmentation. $L_{ce}(s_f, \phi(s_m))$ is the classical categorical cross-entropy defined as

$$L_{ce}(s_f, \phi(s_m)) = \frac{1}{\Omega} \sum_{x \in \Omega} ce(onehot(s_m^w(x)), onehot(s_f(x))), \tag{12}$$

where $s_m^w = \phi(s_m)$, x is the pixel index, $onehot()$ is the one-hot coding function, and $ce()$ is the cross-entropy. $L_{ae}(s_f, \phi(s_m))$ is the squared Euclidean distance between the low-dimensional representation vector of the segmentations after the encoder in Fig. 2, namely,

$$L_{ae}(s_f, \phi(s_m)) = ||encoder(\phi(s_m)) - encoder(s_f)||_2^2. \tag{13}$$

The total loss function is therefore defined as

$$L(I_f, \phi(I_m)) = -L_{img} + L_{as} + \lambda_{df} \; L_{df}, \tag{14}$$

where L_{df} is the field smoothness constraint defined as the total variation of the displacement field. In (14), λ_{ncc}, λ_{ssim}, λ_{ae}, λ_{ce} and λ_{df} are the weight parameters.

4 Experiments

In this section, to demonstrate the effectiveness of the proposed method, we carry out the experiments on public datasets and compare its performance to the state-of-the-art methods, which are two traditional methods SimpleElastic [27] and SyN [2], and three deep CNN based methods, the baseline AC-RegNet [26], GRNet [25], and RCINet [49].

4.1 Image datasets and evaluation metrics

The experiments are conducted in the context of inter-subject 2D chest X-ray image registration, which is quite challenging due to the large anatomical variability between different subjects. Thanks for the preprocessing work by Lucas et al. [26], we use their released datasets as follows

1. Japanese Standard Digital Image Database (JSRT) [32]: it contains 247 images with ground truth segmentation labels. 197 randomly selected samples are used for training and the rest 50 for testing.
2. Montgomery County X-ray Database (MONT) [8]: it contains 138 images with ground truth labels. 110 randomly selected images are used for training and the rest 28 for testing.
3. Shenzhen Hospital X-ray Database (SHEN) [21]: it contains 550 images with ground truth labels. Randomly selected 440 samples are used for training and the rest for testing.

All images from the three datasets have two sizes of 256 × 256 and 64 × 64. Images of 64 × 64 are used for training and 256 × 256 for testing. In this work, we remove the heart parts in JSRT images so that all images have only lungs. Different from the work in [25, 26] where the testing is conducted on the 200 random pairs formed from the test list, we test all pairs of the test set.

We evaluate the methods from two perspectives, the agreement between the warped moving segmentation mask and the fixed and the quality of the displacement field itself. The three segmentation similarity metrics are 1) Dice Similarity Coefficient (DSC), which measures the overlapping between the segmentations [12], 2) Hausdorff Distance (HD), which is the maximum distance between segmentation contours, and 3) Average Symmetric Surface Distance (ASSD), which is the average distance between the segmentation contours. DSC value varies between 0 and 1. HD and ASSD distance have a unit of millimeter. The higher the DSC value, or the smaller HD or ASSD value, the better the registration is. As an indicator of the field smoothness, the Jacobian folding coefficient (JAC) [35] calculates the number of folded pixels in the displacement field.

4.2 Implementation details

We train the model of AC-RegNet and GRNet several times using the released code and parameter settings to get the best possible model for performance comparison. Before training our registration model, we first train the attention model and the anatomical auto-encoder. The weight factor λ in (3) is set to 1.75 and the λ in (4) is 2.0. To train the registration network, the weight factors for loss function (14) are $\lambda_{ncc} = 1.0$, $\lambda_{ssim} = 1.0$, $\lambda_{ae} = 0.1$, $\lambda_{ce} = 1.0$ and $\lambda_{df} = 3.5$. With the parameter settings, we train the three models for each dataset.

4.3 Results

Table 1 shows the registration performance comparison among the traditional methods SimpleElastic [27], SyN [2], deep learning methods RCINet [49], GRNet [25], and the baseline AC-RegNet [26] in terms of mean DSC, HD, ASSD, and JAC scores. It should be noticed that the result of RCINet in the first test is marked with an asterisk sign as the images used in the original test contain the heart parts, while for the rest methods, the heart parts are removed. We still put the result of RCINet for a rough comparison.

From Table 1, we can see that traditional methods SimpleElastic and SyN have comparable performance, and our method is consistently better than AC-RegNet, RCINet, and GRNet in DSC, HD, and ASSD scores. In DSC score, the improvement over AC-RegNet are $0.970 - 0.953 = 0.017$, $0.963 - 0.946 = 0.017$ and $0.957 - 0.929 = 0.028$ for dataset JSRT, MONT and SHEN respectively. The improvement over RCINet and GRNet is about one percentage point. In score HD and ASSD, our method is also significantly better than AC-RegNet, RCINet, and GRNet, except that the HD score is comparable to that of GRNet on the MONT dataset (13.283 vs. 13.385). From the perspective of displacement field quality, the JAC scores of our method are substantially better than that of AC-RegNet, RCINet, and GRNet with a large margin except that the JAC score of RCINet on MONT is 3. Figure 5 shows several examples for visual comparison of registration effect, where the first two columns are the moving images and their labels, the last two columns are the fixed, column 3 and 4 are the warped moving images by GRNet [25], and column 5 and 6 are the warped result by our method. Dice scores between the fixed and the warped moving label are also shown below

Table 1 Registration performance comparison in terms of mean DSC, HD, ASSD, and JAC scores on three datasets among traditional SimpleElastic [27] and SyN [2], deep learning methods AC-RegNet [26], RCINet [49] and GRNet [25]

Dataset	Method	DSC	HD (mm)	ASSD (mm)	JAC
JSRT	SimpleElastix	0.915 (0.045)	31.408 (15.783)	3.860 (2.071)	#
	SyN	0.895 (0.058)	26.545 (15.617)	4.468 (2.654)	#
	AC-RegNet	0.953 (0.012)	22.367 (15.382)	2.195 (0.845)	740.025 (993.758)
	RCINet*	0.950 (0.017)	15.865 (7.254)	2.963 (1.106)	329.000 (417.000)
	GRNet	0.960 (0.009)	13.432 (8.667)	1.705 (0.447)	508.900 (713.700)
	Ours	0.970 (0.005)	9.230 (4.760)	1.244 (0.237)	48.236 (149.108)
MONT	SimpleElastix	0.888 (0.086)	31.097 (22.229)	4.301 (3.743)	#
	SyN	0.842 (0.113)	28.221 (17.701)	5.549 (4.578)	#
	AC-RegNet	0.946 (0.025)	14.841 (16.170)	2.645 (1.768)	692.614 (1074.659)
	RCINet	0.957 (0.031)	19.698(25.684)	2.842 (3.229)	3.000 (17.000)
	GRNet	0.952 (0.024)	13.283 (16.070)	1.810 (1.653)	424.700 (820.100)
	Ours	0.963 (0.016)	13.385 (16.000)	1.722 (1.600)	263.380 (710.760)
SHEN	SimpleElastix	0.908 (0.058)	29.133 (19.261)	3.666 (2.591)	#
	SyN	0.863 (0.087)	32.915 (19.622)	5.214 (3.487)	#
	AC-RegNet	0.929 (0.029)	24.518 (16.248)	2.882 (1.396)	887.145 (1080.631)
	RCINet	0.942 (0.023)	167.227 (119.669)	26.363 (11.511)	150.000 (312.000)
	GRNet	0.943 (0.025)	16.854 (13.616)	2.196 (1.055)	417.400 (750.600)
	Ours	0.957 (0.011)	11.720 (9.376)	1.637 (0.536)	57.464 (203.295)

In the first test, the results of RCINet* are slightly different from the rest as the heart parts in the images are not removed. The numbers in the parenthesis are the standard deviation

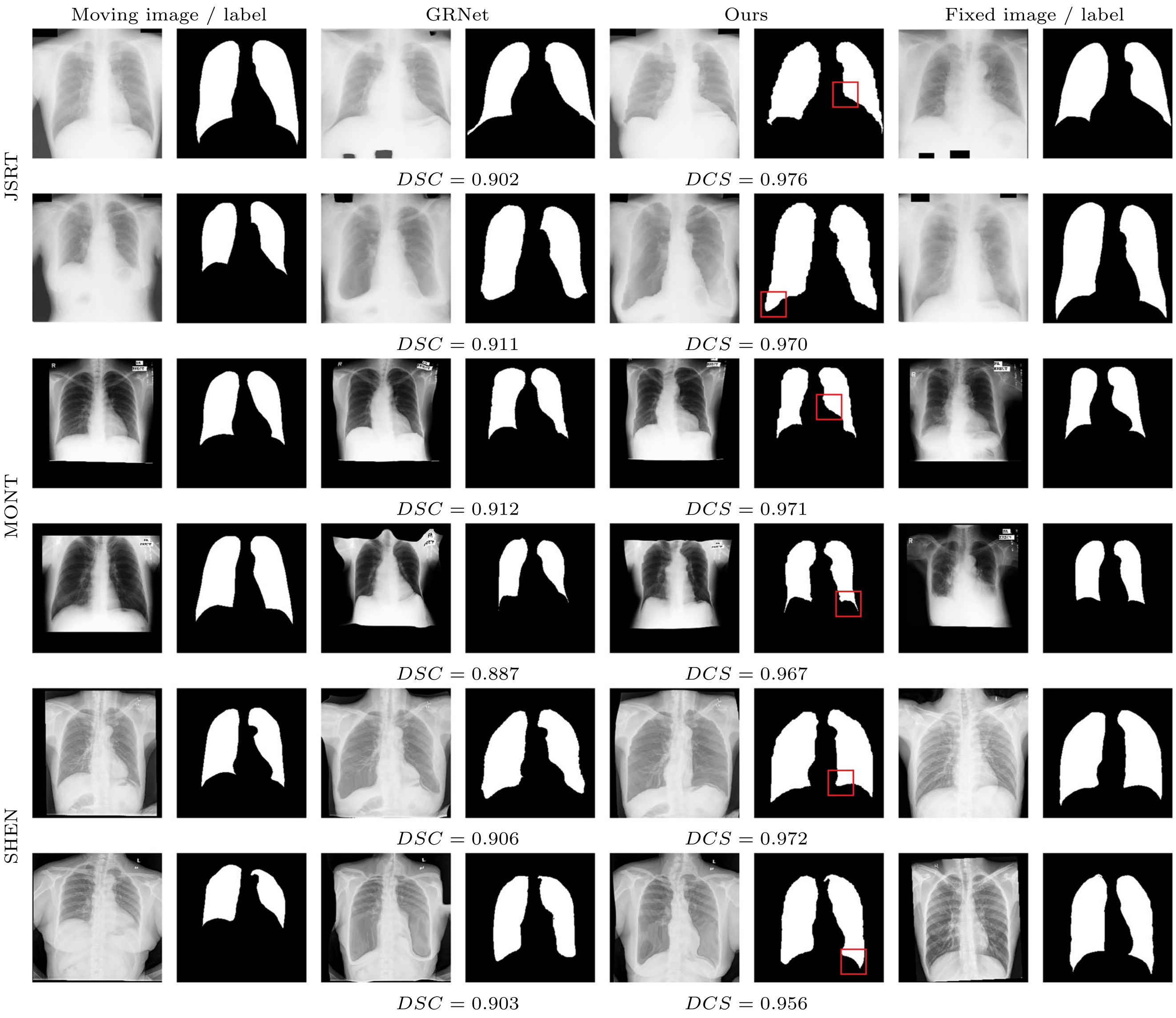


Fig. 5 Example of image registrations. The first two columns are the moving images and their labels. The last two columns are the fixed. Column 3 and 4 are the warped moving by GRNet [25] and column 5 and 6 are the warped result by our proposed method. Dice scores between the fixed and the warped moving label by the two methods are also shown below the image.s

the images. We can see that the global shape or contour of the warped moving label images produced by our method are closer to the fixed.

4.4 Ablation studies

In this section, we examine the contribution of different components of the proposed method.

4.4.1 Effect of contour loss constraint

In our method, we propose contour loss in (2) to improve the performance of the attention segmentation network and the anatomical segmentation auto-encoder. This loss term aims to enhance the global shape of the predicted segmentation mask. Table 2 shows that adding the contour loss to the original slightly improves the prediction accuracy of the attention segmentation network. Figure 6 shows several examples, where the first column is the input images, the second column is the original segmentation using cross-entropy only, the third column is the prediction with the

Table 2 Improvement of attention segmentation with the contour loss

	JSRT	MONT	SHEN
Cross-entropy only	0.961	0.965	0.951
Add contour loss	0.972	0.971	0.962

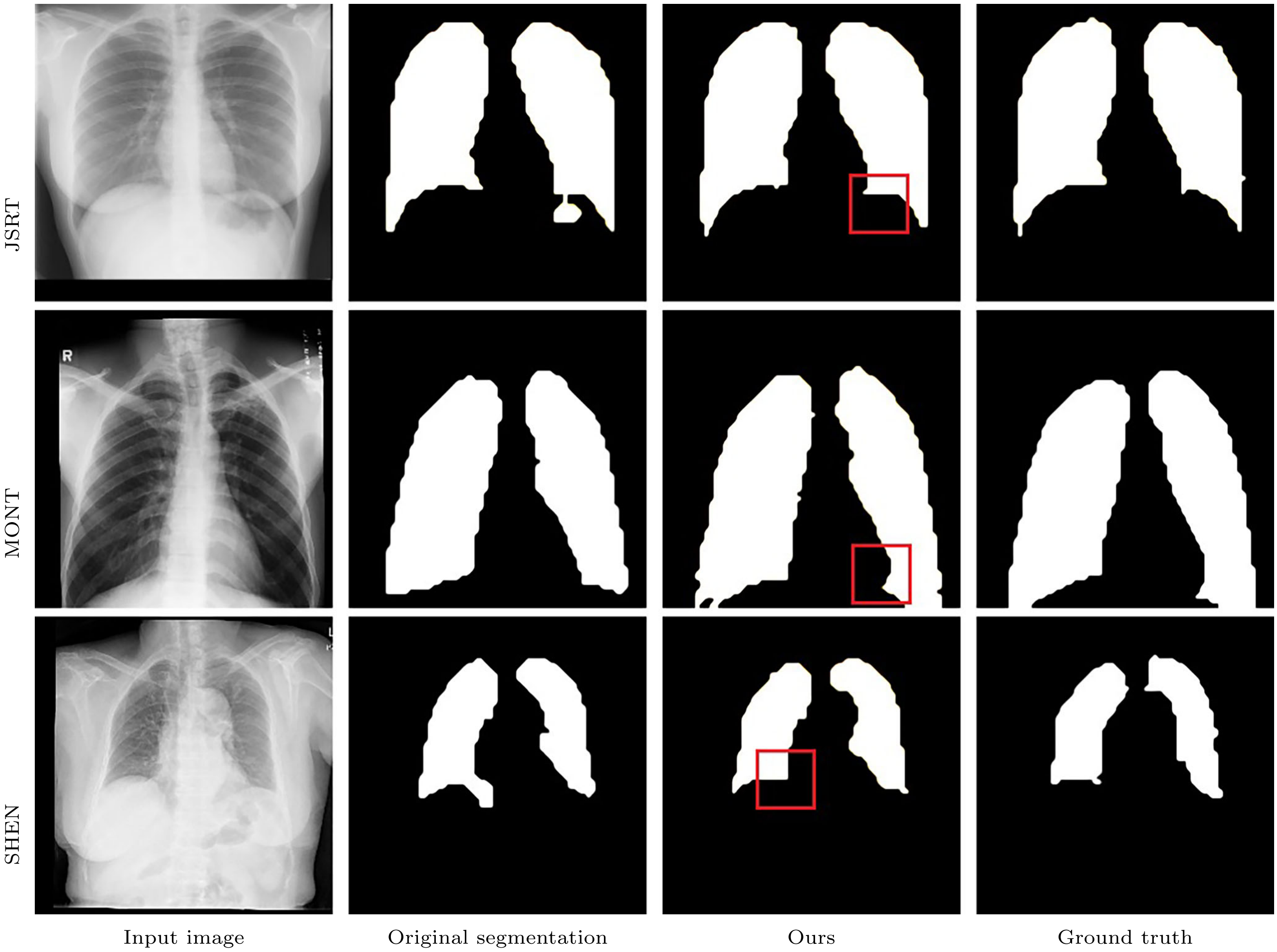


Fig. 6 Example of improved attention segmentation with the contour loss. The first column is the input image, the second column is the original segmentation, the third column is the prediction with the contour loss, and the last column is the ground truth

contour loss, and the last is the ground truth. With the contour loss, the predicted segmentation mask has a better global shape. Similarly, Table 3 and Fig. 7 compares the reconstruction of the decoder when the contour loss is added in the training process. In Table 3, the DSC score is comparable on JSRT and SHEN, and significantly improved on MONT dataset. In terms of HD and ASSD scores, the improvement is very substantial, which means that the global shape of the reconstructed mask becomes much better.

4.4.2 Effect of SSIM loss constraint

The SSIM loss term in (14) is proposed to improve the structural similarity of the aligned image after registration. Figure 8 shows the improvement of the average SSIM score when SSIM loss is used in the training process. The SSIM gain is 0.02, 0.013, and 0.01 for the three datasets, respectively. It should be noted that the SSIM scores are calculated with the cleaned input images where the attention mask removes the regions of the uninterested.

4.4.3 Effect of attention segmentation

Table 4 compares the registration performance with or without the proposed attention segmentation, which aims to

Table 3 Improved reconstruction of the anatomical auto-encoder with the contour loss

Dataset	Method	DSC	HD	ASSD
JSRT	Original AC-RegNet	0.955	22.607	2.794
	With contour loss	0.956	15.686	2.652
MONT	Original AC-RegNet	0.928	38.978	4.718
	With contour loss	0.950	18.161	2.962
SHEN	Original AC-RegNet	0.962	101.989	17.085
	With contour loss	0.968	93.529	14.716

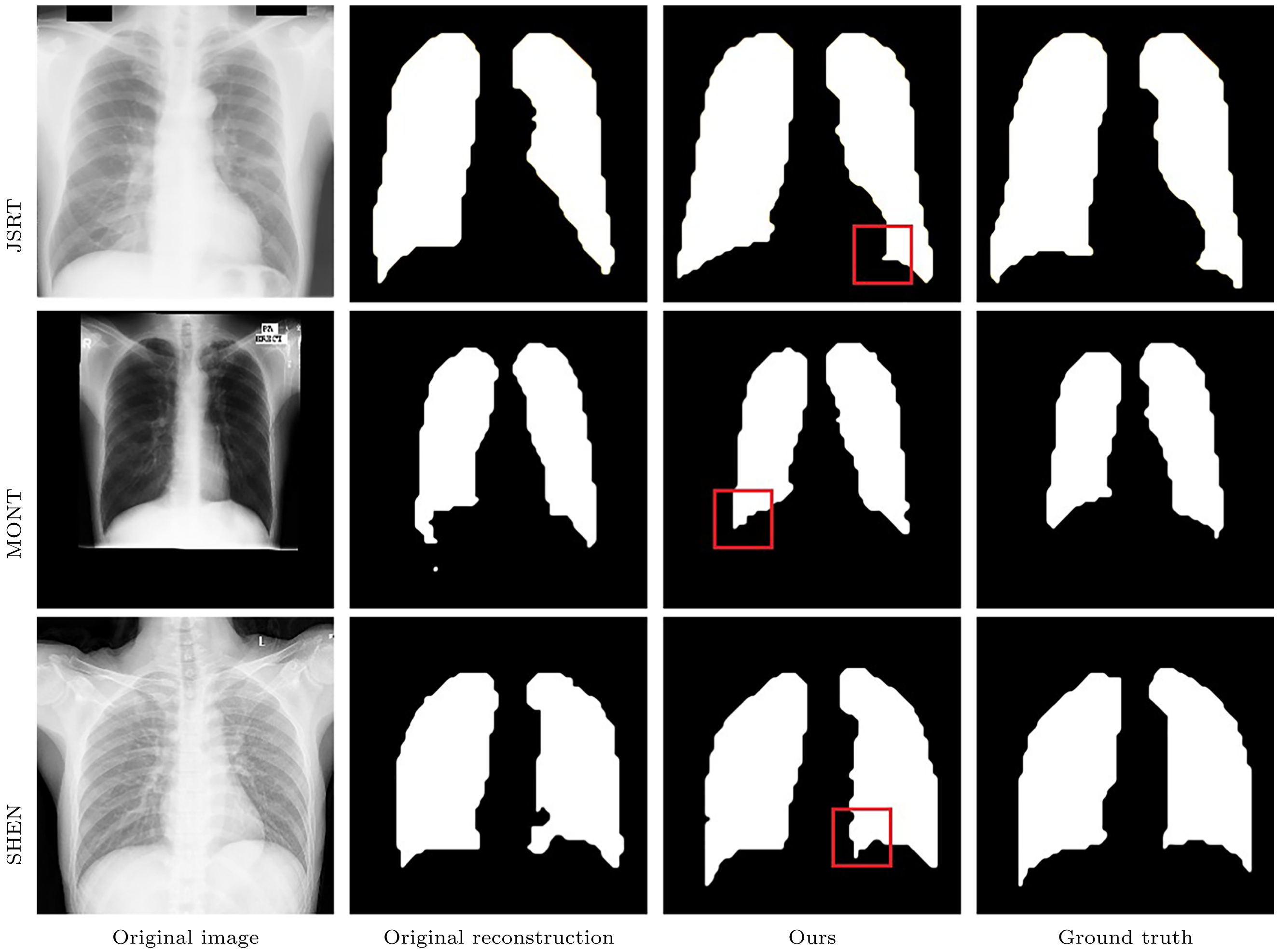


Fig. 7 Effect of contour loss on anatomical encoder-decoder. The first column is the input image, the second column is the original reconstructed label image, the third column is the prediction with the contour loss, and the last column is the ground truth

remove the interference from the uninterested areas in the input images. We can see from Table 4 that using attention segmentation improves the registration accuracy. Take the DSC score for instance, using the attention segmentation, the average DSC improves $0.970 - 0.957 = 0.013$, $0.963 - 0.943 = 0.02$, and $0.957 - 0.935 = 0.022$ on the JSRT, MONT, and SHEN datasets respectively. For HD, ASSD, and JAC scores, the improvement is much more significant.

4.4.4 Effect of multi-scale displacement field fusion

In our proposed method, the dense registration network combines the displacement fields of multiple scales to improve the registration accuracy, especially for those images with large deformations. Figure 9 shows the box plots of the DSC, HD, ASSD, and JAC scores when the number of displacement field scales increases from 1 to 4, in which the asterisks are the mean values. From the figure, we see that on all three datasets, the DSC increases, and HD, ASSD, and JAC decrease when more and more displacement fields of different scales are combined. Especially, the average JAC score shows a consistent trend

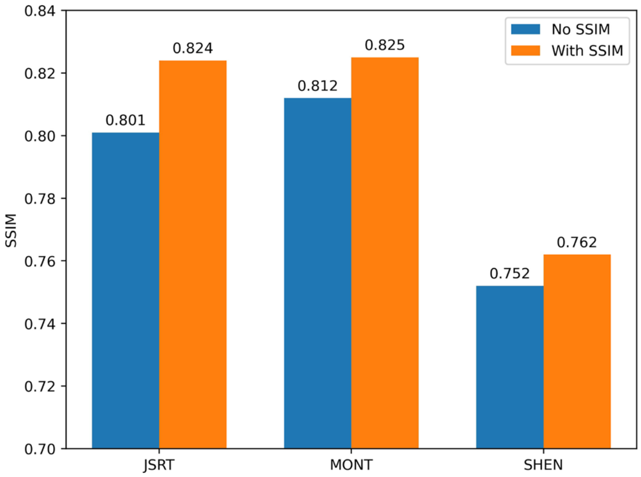


Fig. 8 Using the SSIM loss term improves the registration model. The scores are calculated over the cleaned input images after the attention segmentation

Table 4 Comparison of registration with or without attention segmentation

Dataset	Method	DSC	HD	ASSD	JAC
JSRT	No attention	0.957 (0.012)	15.010 (15.010)	1.823 (0.472)	125.859 (247.386)
	With attention	0.970 (0.005)	9.230 (4.760)	1.244 (0.237)	48.236 (149.108)
MONT	No attention	0.943 (0.031)	17.235 (16.746)	2.183 (1.886)	658.791 (856.494)
	With attention	0.963 (0.016)	13.385 (16.000)	1.722 (1.600)	263.380 (710.760)
SHEN	No attention	0.935 (0.028)	18.518 (13.408)	2.483 (1.141)	63.923 (213.706)
	With attention	0.957 (0.011)	11.720 (9.376)	1.637 (0.536)	57.464 (203.295)

Using the attention segmentation improves the registration performance

with the increase of the number of scales. Figure 10 shows several examples, where the first and last two columns are the moving and fixed images and their labels, the third column is the result when only the topmost branch in Fig. 3 is used for registration. The fourth column is the result when all four scales of displacement fields are combined. Visually, we can easily see the improvement in the fourth column when compared to the third column.

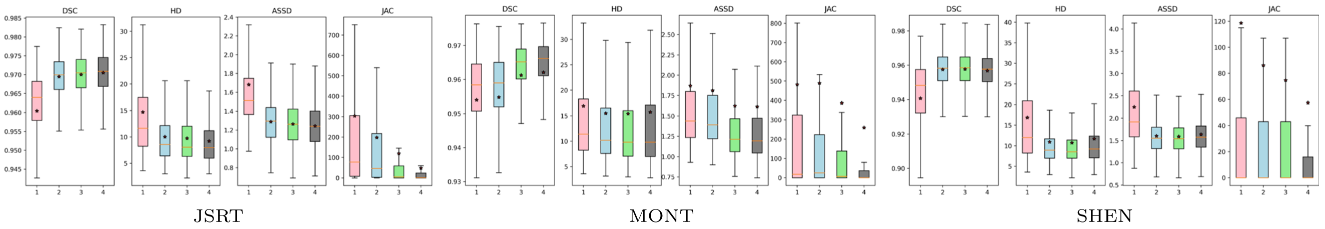
5 Discussion

The major contributions of our proposed method are 1) the hard attention segmentation to remove the interference from the uninterested image areas, 2) the dense registration network based on the weighted fusion of multi-scale displacement fields, 3) and the loss terms to regularize the training process for better model learning.

From Table 2 we see that the segmentation prediction accuracy is relatively high even without our contour loss, which is the reason that we choose to predict the attention segmentation mask using a separate model. The second reason for this external segmentation is that we prefer to make the main registration network not too complicated. Adding the contour loss improves the attention segmentation, especially it helps improve the global shape and reduce the isolated points as shown in Fig. 6. This attention segmentation could be further improved if we use more advanced segmentation networks. However, that might increase the computation cost.

The dense registration network has a U-Net structure that can combine the displacement fields of different scales or resolutions. The maximum number of scales is set to 4 in this work due to the image size of the last scale, i.e., the fourth branch in Fig. 3. The fusion of multi-scale displacement fields makes a significant difference as shown in Figs. 9 and 10.

To improve the training performance, we propose two new loss terms to regularize the model learning, the contour loss of (2) and the SSIM loss of (10). The contour loss is mainly used to train the attention segmentation model and the anatomical auto-encoder, and the SSIM constraint is used in the training of the dense registration model. From the ablation studies, we see the contribution of these loss constraints. One issue with the multi-term loss function is the tuning of the weight factors, see (14). We empirically tune the weight factors on the training dataset as grid search for the best configuration requires a large number of model training. We expect to see more improvement if better tuning of these weight factors is carried out.



JSRT MONT SHEN

Fig. 9 Average registration accuracy with respect to the number of displacement field scales that are combined. It shows that the registration performance increases with more scales of displacement fields combined. On all three datasets, the DSC increases, and HD, ASSD, and JAC decrease. Especially, the average JAC score shows a consistent trend with the increase of the number of scales. The X-axis is the number of displacement field scales from top to down in Fig. 3

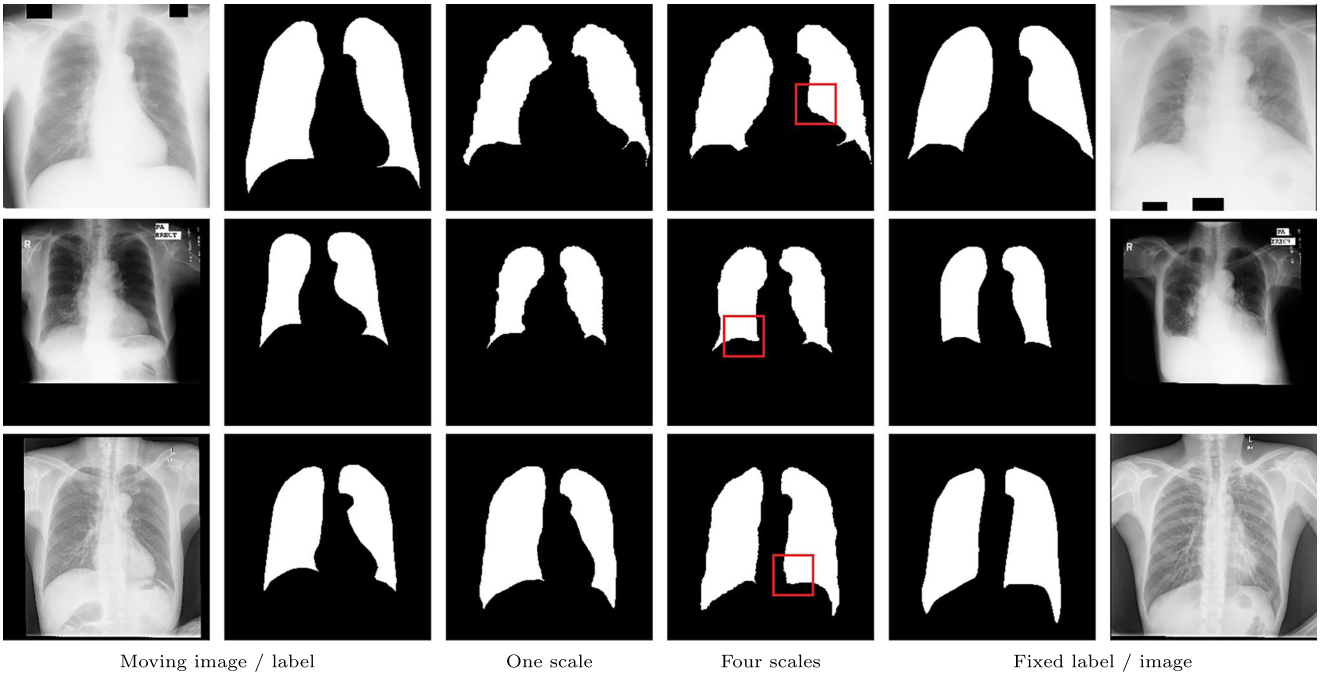


Fig. 10 Effect of multi-scale displacement fields fusion. The first two columns are the moving images and ground truth labels. The last two columns are the fixed. The third column is the warped labels when only one scale, i.e., the topmost branch in Fig. 3 is used. The fourth column is the result when the displacement fields from all branches are combined. It is evident that combining multi-scale displacement fields yields better registration results

6 Conclusion

In this work, we have proposed a novel deformable image registration method based on the attention-guided fusion of multi-scale displacement fields to improve the image registration performance, especially for images with large deformations. Specifically, we propose to adopt a separately trained segmentation network to segment the region of interest, aiming to remove the interference from the uninterested areas in the image. We design a dense registration network that can combine the displacement fields of different scales using learned attention weights for final registration. To improve the registration performance further, we propose a contour loss and image structural similarity based loss (SSIM) to regularize the model learning. Our experimental results on three benchmark datasets have shown significant improvement in DSC, HD, ASSD, and JAC metrics when compared to the state-of-the-art methods. Our method can be directly used in practical medical image registration used in applications ranging from computer assisted diagnosis to computer aided therapy and surgery. In our future work, we shall explore more options to improve the registration performance, such as predicting the velocity field in a diffeomorphic manner instead of the direct displacement field and designing a more advanced deep neural network to further improve the quality of the predicted fields. We will also plan to investigate the extension of the method for multi-modal medical images registration, such as MRT-CT and 2D-3D.

Acknowledgments This work is supported by the National Natural Science Foundation of China under grants 61971290, 61771322, 61871186, and the Fundamental Research Foundation of Shenzhen under Grant JCYJ20190808160815125.

Declarations

Conflict of Interests 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.

References

1. Ashburner J (2007) A fast diffeomorphic image registration algorithm. Neuroimage 38(1):95–113
2. Avants BB, Epstein CL, Grossman M, Gee JC (2008) Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12(1):26–41
3. Avants BB, Tustison N, Song G (2009) Advanced normalization tools (ants). Insight j 2(365):1–35
4. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV (2018) An unsupervised learning model for deformable medical image registration. IEEE
5. Balakrishnan G, Zhao A, Sabuncu MRS, Guttag J, Dalca AV (2019) Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans Med Imaging, 1788–1800
6. Beg MF, Miller MI, Trouvé A, Younes L (2005) Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int J Comput Vision 61(2):139–157
7. Beg MF, Miller MI, Trouvé A, Younes L (2005) Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int J Comput Vision 61(2):139–157
8. Candemir S, Jaeger S, Palaniappan K, Musco JP, Singh RK, Xue Z, Karargyris A, Antani SK, Thoma GR, McDonald CJ (2014) Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans Medical Imaging 33(2):577–590
9. Cao X, Yang J, Zhang J, Nie D, Kim M, Wang Q, Shen D (2017) Deformable image registration based on similarity-steered cnn regression. Springer Cham
10. Ceritoglu C, Oishi K, Li X, Chou MC, Younes L, Albert M, Lyketsos C, van Zijl PC, Miller MI, Mori S (2009) Multi-contrast large deformation diffeomorphic metric mapping for diffusion tensor imaging. Neuroimage 47(2):618–627
11. Davatzikos C (1997) Spatial transformation and registration of brain images using elastically deformable models. Comput Vis Image Underst 66(2):207–222
12. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3)
13. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: 9Th international conference on learning representations, ICLR 2021, virtual event, austria, may 3-7, 2021. Openreview.net
14. El-Gamal FEZA, Elmogy M, Atwan A (2016) Current trends in medical image registration and fusion. Egypt Inform J 17(1):99–124
15. Estienne T, Vakalopoulou M, Christodoulidis S, Battistela E, Deutsch E (2019) U-reSNet: Ultimate coupling of registration and segmentation with deep nets. Medical image computing and computer assisted intervention – MICCAI 2019, 22nd international conference, Shenzhen, China October 13–17, 2019, Proceedings, Part III
16. Fischer P, Dosovitskiy A, Ilg E, Husser P, Hazrba C, Golkov V, Patrick VDS, Cremers D, Brox T (2016) Flownet: Learning optical flow with convolutional networks. In: 2015 IEEE International conference on computer vision (ICCV)
17. Gaser C (2016) Structural MRI: Morphometry neuroeconomics
18. Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023
19. Hu Y, Modat M, Gibson E, Li W, Ghavami N, Bonmati E, Wang G, Bandula S, Moore CM, Emberton M et al (2018) Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal 49:1–13
20. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K (2015) Spatial transformer networks. arXiv:1506.02025
21. Jaeger S, Karargyris A, Candemir S, Folio LR, Siegelman J, Callaghan FM, Xue Z, Palaniappan K, Singh RK, Antani SK, Thoma GR, Wang Y, Lu P, McDonald CJ (2014) Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging 33(2):233–245
22. Kahaki SM, Wang SL, Stepanyants A (2019) Accurate registration of in vivo time-lapse images. In: Medical imaging 2019: Image processing, vol 10949. International Society for Optics and Photonics, p 109491d
23. Li H, Fan Y (2018) Non-rigid image registration using self-supervised fully convolutional networks without training data. In: 2018 IEEE 15Th international symposium on biomedical imaging (ISBI 2018)
24. Liu X, Li M, Wang L, Dou Y, Yin J, Zhu E (2017) Multiple kernel k-means with incomplete kernels. In: Singh SP, Markovitch S (eds) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. AAAI Press, pp 2259–2265
25. Luo Y, Cao W, He Z, Zou W, He Z (2021) Deformable adversarial registration network with multiple loss constraints. Comput Med Imaging Graph 91(6):101931
26. Mansilla L, Milone DH, Ferrante E (2020) Learning deformable registration of medical images with anatomical constraints. Neural Netw 124:269–279
27. Marstal K, Berendsen F, Staring M, Klein S (2016) Simpleelastix: a user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 134–142
28. Modersitzki J (2008) Fair: Flexible algorithms for image registration. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718843
29. Rohé MM, Datar M, Heimann T, Sermesant M, Pennec X (2017) Svf-net: Learning deformable image registration using shape matching. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 266–274
30. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. arXiv:1505.04597
31. Shan S, Yan W, Guo X, Chang IC, Fan Y, Xu Y (2017) Unsupervised end-to-end learning for deformable medical image registration
32. Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu KI, Matsui M, Fujita H, Kodera Y, Doi K (2000) Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am J Roentgenol 174(1):71–74
33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December 4-9, 2017 Long Beach, CA, USA, pp 5998–6008
34. Vercauteren T, Pennec X, Perchant A, Ayache N (2009) Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage 45(1):S61–S72

35. Vos D, Bob D, Berendsen FF, Viergever MA, Sokooti H, Staring M (2018) A deep learning framework for unsupervised affine and deformable image registration. Medical Image Analysis
36. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification. In: 2017 IEEE Conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, pp 6450–6458
37. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q (2020) Eca-net: Efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on computer vision and pattern recognition, CVPR 2020, seattle, WA, USA, June 13-19, 2020. IEEE, pp 11531–11539
38. Wang X (2007) Laplacian operator-based edge detectors. IEEE Trans Pattern Anal Mach Intell 29(5):886–90
39. Wang X, Girshick RB, Gupta A, He K (2018) Non-local neural networks. In: 2018 IEEE Conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. IEEE Computer Society, pp 7794–7803
40. Wang Z (2004) Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing
41. Wang Z, Delingette H (2021) Attention for image registration (air): an unsupervised transformer approach. arXiv:2105.02282
42. Woo S, Park J, Lee J, Kweon IS (2018) CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, Lecture Notes in Computer Science, vol 11211. Springer, pp 3–19
43. Wu H, Zhao H, Zhang M (2021) Not all attention is all you need. arXiv:2104.04692
44. Xiao Y, Zhou Z (2020) Infrared image extraction algorithm based on adaptive growth immune field. Neural Process Lett 51(3):2575–2587
45. Yu X, Ye X, Gao Q (2019) Pipeline image segmentation algorithm and heat loss calculation based on gene-regulated apoptosis mechanism. Int J Press Vessel Pip 172:329–336
46. Zhao H, Jia J, Koltun V (2020) Exploring self-attention for image recognition. In: 2020 IEEE/CVF Conference on computer vision and pattern recognition, CVPR 2020, seattle, WA, USA, June 13-19, 2020. IEEE, pp 10073–10082
47. Zhao S, Dong Y, Chang EI, Xu Y et al (2019) Recursive cascaded networks for unsupervised medical image registration. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10600–10610
48. Zhao S, Lau T, Luo J, Eric I, Chang C, Xu Y (2019) Unsupervised 3d end-to-end medical image registration with volume tweening network. IEEE J Biomed Health Inform 24(5):1394–1404
49. Zou W, Luo Y, Cao W, He Z, He Z (2021) A cascaded registration network rcinet with segmentation mask. Neural Computing and Applications

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

Zhiquan He is currently working as assistant professor in College of Information Engineering, Shenzhen University, China. He received his M.S. degree from Institute of Electronics, Chinese Academy of Sciences in 2001, and the PhD degree from the department of Computer Science, University of Missouri-Columbia in 2014. His research area is in the areas of image processing, computer vision and machine learning.

Yupeng He is currently pursuing the M.Eng. degree in Electronic Information in Shenzhen University, Shenzhen, China. His research interests are image processing and Artificial intelligence.

Wenming Cao received the M.S. degree from the System Science Institute, China Science Academy, Beijing, China, in 1991, and the Ph.D. degree from the School of Automation, Southeast University, Nanjing, China, in 2003. From 2005 to 2007, he was a postdoctoral Researcher with the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with Shenzhen University, Shenzhen, China. He has authored or coauthored more than 80 publications in top-tier conferences and journals. His research interests include pattern recognition, image processing, and visual tracking.