Biomedical Signal Processing and Control 81 (2023) 104403

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Similarity attention-based CNN for robust 3D medical image registration

Fei Zhu [a,b], Sheng Wang [a,b], Dun Li [c], Qiang Li [a,b,*]

[a] Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, Hubei, China
[b] MoE Key Laboratory for Biomedical Photonics, Collaborative Innovation Center for Biomedical Engineering, School of Engineering Sciences, Huazhong University of Science and Technology, Wuhan, Hubei, China
[c] United Imaging Surgical Healthcare Co., Ltd., China

ARTICLE INFO

Keywords:
Convolutional neural network
Medical image registration
Similarity
Attention
Multi-scale

ABSTRACT

In recent years, deep learning (DL)-based registration technology has significantly improved the calculation speed of medical image registration. Existing DL-based registration methods generally use raw data features to predict the deformation field. However, this strategy may not be very effective for difficult registration tasks. Hence, in this study, we propose a similarity attention-based convolutional neural network (CNN) for accurate and robust three-dimensional medical image registration. We first introduce a similarity-based local attention model as an auxiliary module for building a displacement searching space, instead of a direct displacement prediction based on raw data. The proposed model can help the network focus on spatial correspondences with high similarities and ignore those with low similarities. A multi-scale CNN is then integrated with the similarity-based local attention for providing non-local attention, lightweight network, and coarse-to-fine registration. We evaluated the proposed method for various applications, such as the registration of large-scope abdominal computerized tomography (CT) images and chest CT images acquired at different respiratory phases, and atlas registration in magnetic resonance imaging. The experimental results demonstrate that the proposed method can provide a more accurate and robust registration performance than state-of-the-art registration methods.

1. Introduction

Medical image registration has extensive applications in clinic, such as image guidance [1–2], motion tracking [3–4], segmentation [5–6], and image reconstruction [7–8]. Fast and accurate automatic registration algorithm is very important for the development of modern medicine.

The objective of medical image registration is to determine the spatial correspondence between a pair of input images (the reference image and the moving image). Over the past few decades, traditional iteration-based registration methods have become the most popular registration techniques [9–10]. These methods essentially consist of three components: the deformation model, objective function, and optimization strategy. Several innovative techniques have been proposed based on these registration methods [11–15] and have achieved satisfactory results on several datasets. However, these iteration-based registration methods are often time-consuming and impractical for nearly real-time clinical applications.

Recently, the emergence of deep learning (DL) has provided the possibility for real-time registration and transformed the landscape of medical image registration research. The current DL-based registration methods for three-dimensional (3D) non-rigid medical image registration primarily include the deep similarity-based registration methods, supervised transformation prediction registration methods, and unsupervised transformation prediction registration methods.

The deep similarity-based registration method uses a DL network to produce a deep similarity metric instead of traditional image similarity measures, such as the sum of squared differences, mean absolute differences, and mutual information (MI) in iteration-based image registration. Simonovsky et al. [16] trained a deep similarity learning network to obtain the similarity metric using a few aligned image pairs and subsequently used the learned deep similarity metric to construct the objective function for a brain T1-T2 registration task. Haskins et al. [17] utilized a convolutional neural network (CNN) to obtain a similarity metric that replaced MI for the rigid registration of 3D magnetic resonance and transrectal ultrasound images. Experimental results

* Corresponding author at: Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, Hubei, China.
E-mail addresses: zhufei@hust.edu.cn (F. Zhu), m202073227@hust.edu.cn (S. Wang), alex.li@united-imaging.com (D. Li), liqiang8@hust.edu.cn (Q. Li).

https://doi.org/10.1016/j.bspc.2022.104403
Received 3 August 2022; Received in revised form 31 October 2022; Accepted 7 November 2022
Available online 22 November 2022

confirmed that, compared with MI and its variant [18], the learned similarity metric can provide better registration results. Although the deep similarity metric outperforms the traditional similarity metrics in medical image registration, the registration process itself still inherits the iterative nature of traditional registration frameworks, which reduces its registration efficiency.

To predict transformation in one-shot, several researchers have attempted to train networks to directly infer the final transformation in a single forward prediction. Initially, known transformations were used to train a supervised DL registration method for producing the deformation field. Sun et al. [19] used expert-aligned computerized tomography and ultrasound (CT-US) image pairs as the ground truth to train the network for simulated CT-US registration. However, it has limited functionality on real CT-US pairs owing to the vast differences in appearance between the simulated and real US images. Eppenhof et al. [20] used synthetic random transformations to obtain the training data, which were subsequently used to train a CNN for 3D CT lung registration. Because the simulated transformation often differs considerably from the true transformation, Sentker et al. [21] used a traditional registration methods, such as PlastiMatch [22], NiftyReg [23], and VarReg [24], to generate the deformation field as the ground truth. They proved that the performance of the developed DL registration method was comparable to that of traditional registration methods on DIRLAB [25] datasets. These supervised transformation prediction registration methods have achieved end-to-end registration, but they are considerably limited in acquiring the ground truth.

Therefore, unsupervised image registration methods have been developed to overcome the lack of training datasets with known transformations. Balakrishnan et al. [26–27] proposed an unsupervised CNN-based registration method, the VoxelMorph, for magnetic resonance imaging (MRI) brain-atlas-based registration. Zhang et al. [28] introduced diffeomorphic mapping in a CNN to avoid the folding of the predicted transformation, and proved that their method outperformed those developed by Demons [11] and Syn [29]. However, these methods may not work well when the image displacement is large or when the image content is complex, which are common in abdominal CT scans. To address these limitations, Zhao et al. proposed an unsupervised affine and deformable image registration framework by stacking multiple CNN into one network, the volume tweening network (VTN) [30], which was later extended to a recursive cascade architecture [31] for CT liver and brain registration. Their network achieved significantly improved registration accuracy compared to state-of-the-art methods on both liver and brain datasets. However, more cascades require more parameters and computation time, and these cascades may exacerbate the folding problem in DL-based registration.

Existing DL-based registration methods typically input the image data, encode the input images to features, and estimate the deformation field based on raw data. However, this method may be too crude to provide good results. Therefore, in this study, we propose a similarity attention-based CNN (SAN) for real-time, accurate and robust 3D medical image registration. To improve the registration performance, a similarity-based local attention model, inspired by the ideas of attention [32] and cost volume [33], is first employed in the proposed network to build a displacement searching space, instead of making a direct prediction based on raw data. This similarity-based local attention model can readily capture the similarity between input images and help the network efficiently learn the spatial correspondence associated with these similarities.

In addition, we designed a multi-scale registration framework to achieve a lightweight network and coarse-to-fine registration. By combining with multi-scale structure, the local similarity-based attention model can span a large image domain at a minimal computational cost; thus, the similarity-based local attention model, which is called the "similarity attention model", becomes non-local.

The key contributions of this study can be summarized as follows:

- A novel similarity-based local attention model is proposed to build a displacement searching space beforehand, instead of making a direct prediction based on raw data in the registration network.
- A multi-scale registration framework combined with the similarity-based local attention model is proposed to achieve non-local attention, lightweight network, and coarse-to-fine registration for robust 3D medical image registration.
- The superior registration performance of the proposed method is presented in extensive experimental results for the registration of large-scope abdominal CT images and chest CT images acquired at different respiratory phases as well as the atlas registration in brain MRI.

2. Related work

2.1. Pyramid DIR network

Recently, several lightweight networks, such as the Laplacian pyramid network (LIPNet) [34] and dual-stream pyramid network (DSPN) [35], have been proposed for 3D medical image registration. These networks decompose the deformation field in a coarse-to-fine manner, wherein a coarse sub-field is predicted by a shallow layer to help estimate a fine sub-field in the next layer. Such coarse-to-fine decomposition can achieve superior registration performance with fewer parameters. However, LIPNet has difficulty obtaining a converged result and must be individually trained, and DSPN performance deteriorates significantly when handling large displacements. Moreover, these two methods rely completely on raw image information and simplistic deformation field estimators; thus, they exhibit poor registration accuracy in some applications.

2.2. Cost volume for medical image registration

Cost volume [36] is a concept used to compute the data-matching costs associated with relating a pixel with its corresponding pixels in the other image. It has unique advantages and is widely used in stereo matching and computing the optical flow [33]. For features $c_1(x_1)$ and $c_2(x_2)$ from two images, the cost volume is defined as

$$cv(x_1, x_2) = \frac{1}{N}(c_1(x_1))^T c_2(x_2) \tag{1}$$

where T is the transpose operator and N is the length of the column vector $c_1(x_1)$.

Recently, He et al. [37] introduced normalized cost correlation volumes (CCV) for the 3D registration of expiratory-inspiratory CT lung images. However, for the processing of cost volumes, the normalization of cost correlation volumes is simplistic and lacks parameters to enhance its intelligence. In addition, the network may not provide sufficiently accurate registration results because the estimator uses as the input only the cost correlation volumes and fixed images. Therefore, this method leaves room for further improvement.

2.3. Attention for medical image registration

The attention mechanism has become an integral part of models for various tasks such as natural language processing and computer vision. Attention can be usually described with the following steps:

Decomposition: mapping a query (Q) and a set of key (K)–value (V) pairs, where the query, key, and value are vectors.
Calculation of the similarity between Q and K.
Normalization of the obtained similarity as an attention map using the softmax operator.
Computation of the weighted sum of V to obtain the output.

Essentially, the attention computes the response at a position in a sequence (e.g., an image) by attending to all other positions. It is also an integral component of transformer [38]. Many researchers have proposed the attention-based registration methods. Table 1 lists 8 state-of-the-art attention-based registration methods and provides brief descriptions of these methods. Most of these methods are based on the transformer approach due to its superiority in feature extraction. However, the transformer-based method usually requires a large amount of training data to ensure good performance, which is often difficult for medical image registration.

In this study, we carefully designed a similarity attention-based DL network for robust 3D medical image registration. Unlike the above methods in which the attention mechanisms or transformers are used to extract features, the similarity attention modular in the proposed method does not directly extract features, but rather processes the extracted features to obtain the similarity between images. Then the similarity relation of the extracted features is used to build a displacement searching space for assisting the image registration, and it can reduce the difficulty of deformation field prediction and thus improve the registration accuracy. In addition, we use the coarse-to-fine global registration strategy to address the challenge of large search range of the non-local image similarity and calculation cost. The coarse-to-fine strategy is a key to make our method robust for the registration of image pairs of large deformation.

3. Methods

The SAN framework comprises an affine registration subnetwork and a deformable registration subnetwork. A diagram of the proposed SAN is shown in Fig. 1. Here, the affine registration subnetwork is necessary before the deformable registration because it can effectively overcome the affine deformation between the input images and reduce the burden on the deformable registration network. The structure of the affine subnetwork is identical to that used by Zhao et al. [31].

Table 1
The descriptions of 8 state-of-the-art attention-based registration methods.

Method	Assumption	Advantage	Limitation
VAN [39]	Voting process for final deformation field improves registration performance	Good feature extraction by channel and spatial attention	Reliance on the accuracy of registration branches
RMAn[40]	Mutual attention and recursive network improve registration performance	Large effective receptive fields and progressive registration	High computing cost and large number of parameters
ViT-V-Net [41]	Vision Transformer improves registration performance	Long-distance relationships between points in images	Poor performance for large deformation
PC-SwinMorph [42]	Patch embeddings are meaningful for performance gain in medical data	Patch-wise contrastive learning for feature extraction	Blocky artefacts and low registration efficiency
C2FViT[43]	Vision Transformer improves registration performance	Large effective receptive fields and progressive registration	Suitable only for affine registration
MS-DIRNet [44]	Attention-gates and adversarial network are important for abdominal 4D-CT registration	Good robustness for large deformation	Blocky artefacts and low registration efficiency
TransMorph [45]	Swin Transformer and uncertainty estimate improve registration performance	Large effective receptive fields and flatter loss landscapes	High computing cost and limited precision for abdominal data
XMorpher [46]	Feature matching during learning improves registration performance	Outstanding performance for registration of fine deformation	High computing cost and poor robustness for large deformation

The deformable registration subnetwork forms the core of the proposed method and determines the final registration accuracy. It consists of three components: a multi-scale feature extractor, similarity attention model, and deformation field estimator. First, the input images are encoded to 3-level scaled feature representations using a multi-scale feature extractor. Next, each scaled feature is used to construct a similarity attention map using the similarity attention model. Finally, the similarity attention map, features of the input images, and the deformation field from the upper scale (there is no deformation field at the bottom level) are used to estimate the dense deformation field using the deformation field estimator. In the following section, we provide a detailed description of the three components.

3.1. Multi-scale feature extractor

The 3-level features for the moving and reference input images after the affine registration are first generated using the multi-scale feature extractor. The multiscale feature extractor includes three encoding blocks and each encoding block consists of a convolution layer and a LeakyReLU [47] activation layer which is used with slope of 0.1. Fig. 2 shows the structure of the multi-scale feature extractor. As shown, all the scaled features are directly derived from the input image rather than from the features in the previous layer, which is different from the existing pyramid-based DL registration methods [34–35]. This feature extraction strategy can reduce the mutual interference of features at different levels and make it easier for the network to train and converge.

3.2. Similarity attention model

In the proposed method, the similarity attention model is the most important part for improving registration performance. The architecture of the similarity attention model is depicted in Fig. 3.

Because the similarity calculation for the entire image is prohibitive in terms of runtime and memory usage, we constructed a local similarity attention map based on multi-scale image features to obtain a non-local similarity response. The local similarity attention for each scaled image feature can be calculated as follows.

Given a patch $P_m(i) \in R^{h\times w\times d\times c}$ centered at i in the features of the moving image and a local search window $P_r(i) \in R^{h\times w\times d\times c}$ centered at i in the features of the reference image, we first transform them into two new feature spaces f and g. Here, h, w, and d are the patch sizes and c is the number of channels. The new feature spaces are obtained as follows:

$$\begin{aligned} P_m^s(i) &= W_m P_m(i) \\ P_r^s(i) &= W_r P_r(i) \end{aligned} \tag{2}$$

Here, W_m and W_r are the learned weight matrices that are implemented as $1 \times 1 \times 1$ convolutions. Subsequently, for the features of the moving image, we apply average pooling to the patch $P_m^s(i) \in R^{h\times w\times d\times c}$ to obtain a two-dimensional (2D) feature matrix $S_m^s(i) \in R^{1\times c}$ representing the coded features at pixel i. For computational convenience, the features $P_r^s(i) \in R^{h\times w\times d\times c}$ in the search window of the reference image are flattened to a 2D feature matrix $S_r^s(i) \in R^{hwd\times c}$. Finally, similarity attention is computed as:

$$SA(i) = S_r^s(i) S_m^s(i)^T \tag{3}$$

where T denotes the transpose operation.

$SA(i)$ stores the similarity relation for associating pixel i with its corresponding pixels in the other image. By applying this calculation to all pixels, the similarity relation of the entire image can be easily constructed. Once the similarity relation matrix is obtained, the modular outputs a similarity attention map similar to the deformation field by

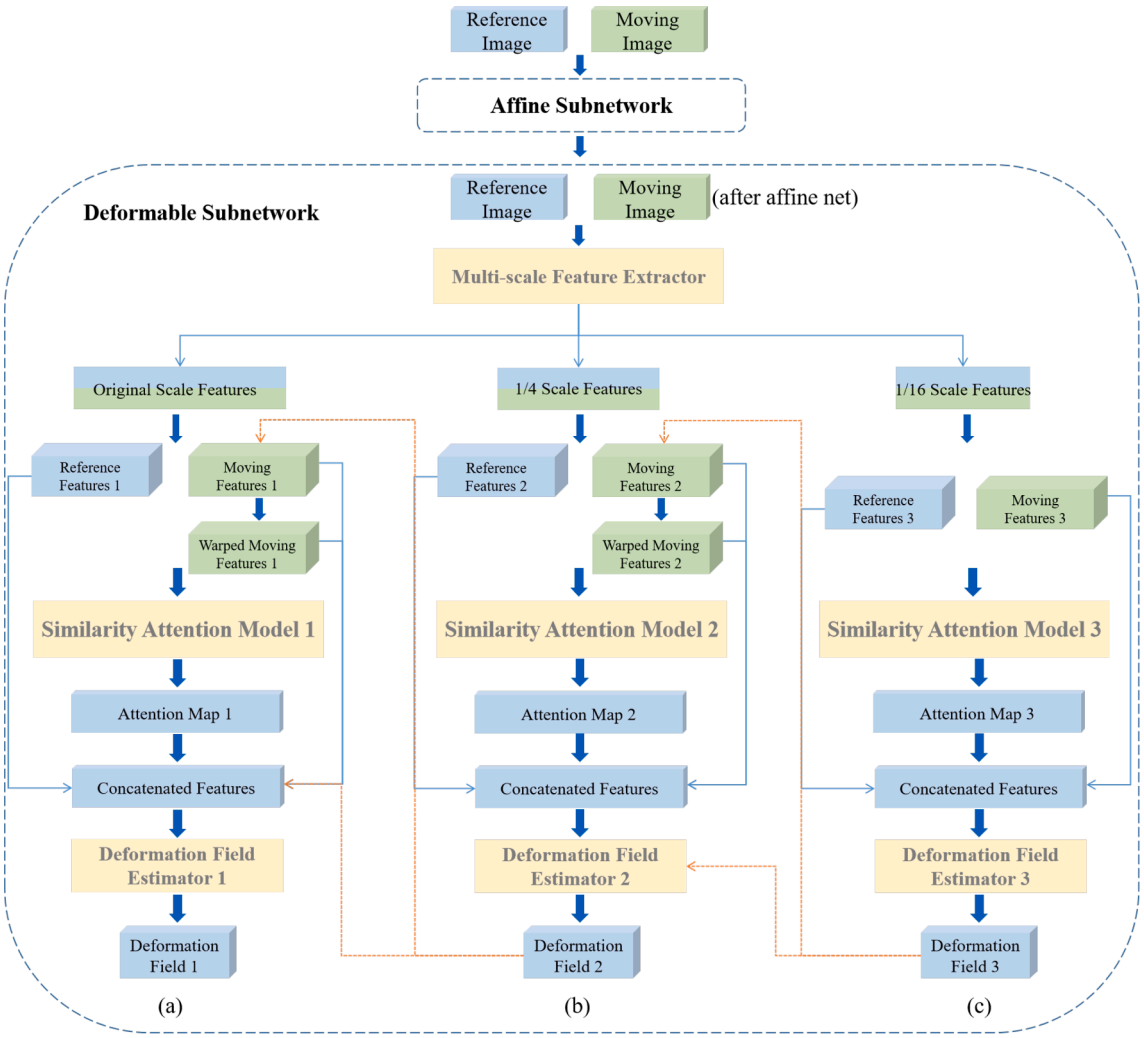

Fig. 1. Framework of SAN. (a) Original scaled level of the deformable registration network. (b) 1/4 scaled level of the deformable registration network. (c) 1/16 scaled level of the deformable registration network.

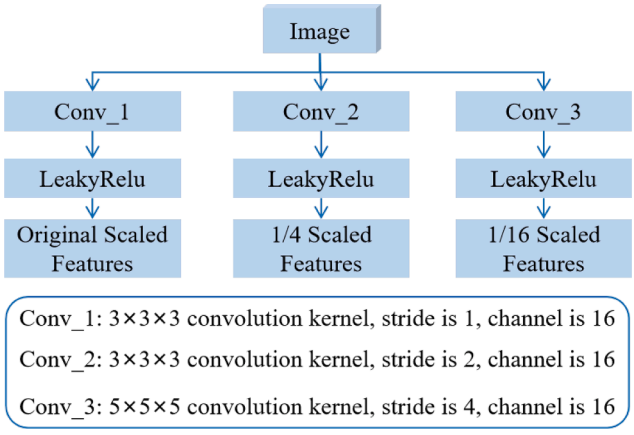

Fig. 2. Structure of the multi-scale feature extractor.

applying a convolution layer to the similarity relation matrix. It should be noted that the similarity attention map at the 1/16 scaled level is computed using the features of the reference image and the moving image. By contrast, the similarity attention maps of the 1/4 scaled and original scaled levels are computed using the features of the reference image and the moving image warped using the 2x upsampled deformation field predicted at the previous level. The warping tool is a spatial transformation network [48].

3.3. *Coarse-to-fine deformation field estimators*

The coarse-to-fine deformation field estimators start by predicting a lower-resolution (coarser) deformation field, wherein neighboring pixels that share the same displacement vector are grouped. Subsequently, they estimate a finer field, wherein fewer pixels are grouped. Finally they yield the deformation field with the same resolution as that of the input image. A diagram of the deformation field estimation process at level i is shown in Fig. 4.

For the 1/16 scaled level, the concatenated features include the reference and moving image features and similarity attention map, which are marked in blue in Fig. 4. In addition, the concatenated features for the 1/4 scaled level and original scaled level include the warped features of the moving image at the current level and the upsampled deformation field from the previous level, which are marked in yellow in Fig. 4. We use several residual network (ResNet) blocks to decode the concatenated features, and the output of a ResNet block serves as input to the subsequent ResNet block. In the experiments, the number of ResNet blocks was set as five (four with 48 feature channels and one with 24).

For the concatenated features, the features of the reference and moving images provide the initial image information. The similarity

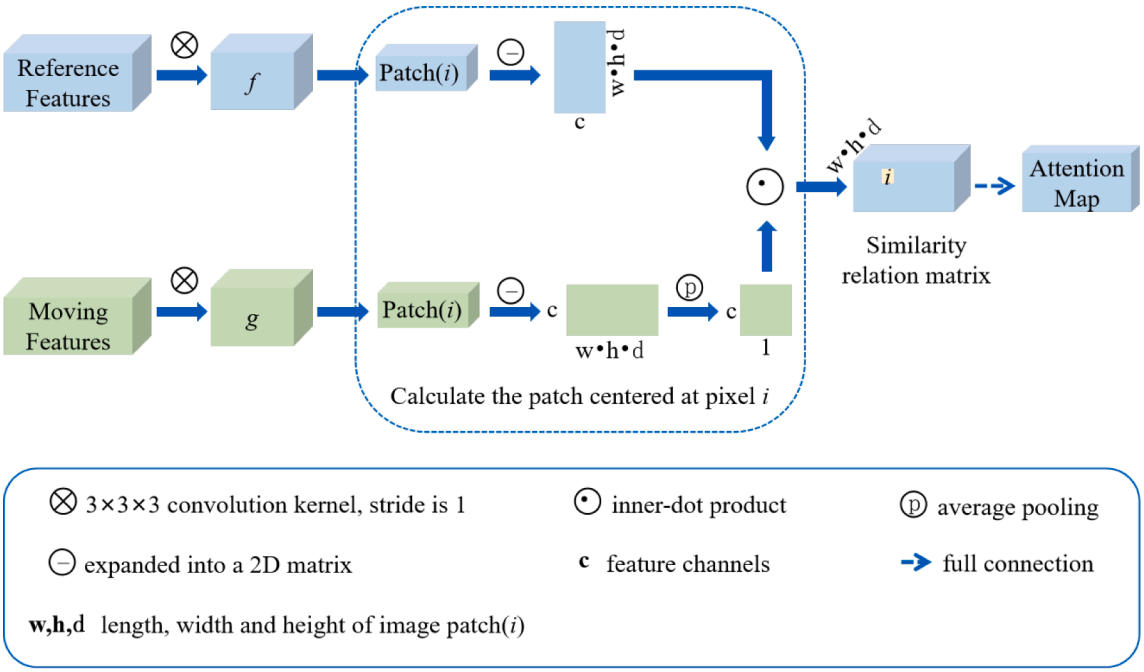

Fig. 3. Architecture of the similarity attention model.

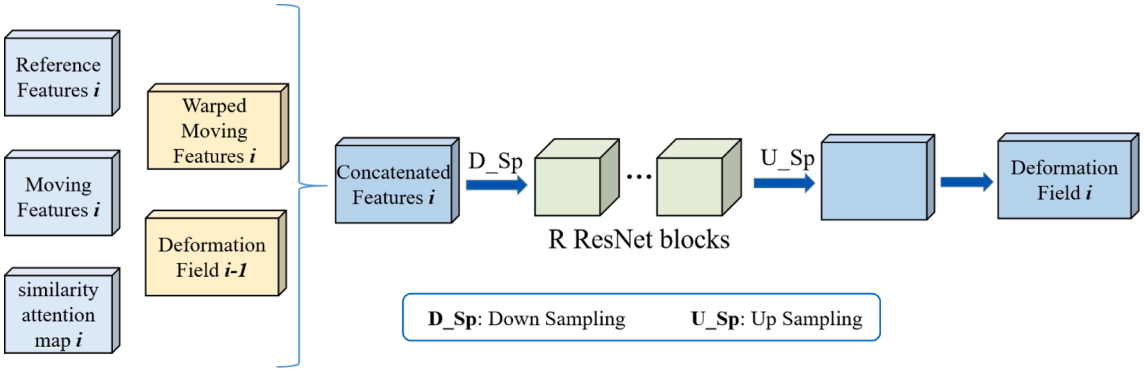

Fig. 4. Structure of the deformation field estimator for level *i*.

attention map and the warped features of the moving image can effectively help the network detect the information match/mismatch. The deformation field refers to the registration information of the previous scale.

3.4. Loss functions

To train the model in an unsupervised manner, two unsupervised losses were employed to measure the (dis)similarity between the reference images and moving images warped by the spatial transformer. For the affine subnetwork, the loss function was the same as that in [31]. For the deformable registration subnetwork, the loss function is designed as follows:

$$L_D = L_{cc} + L_{ncc} + L_{reg} \tag{3}$$

where L_{cc} and L_{ncc} are the correlation coefficient and the local normalized cross-correlation, respectively. Given two images $f, m \in \Omega$, L_{cc} and L_{ncc} are computed as follows:

$$L_{cc} = 1 - \frac{Cov[f,m]}{\sqrt{Cov[f,f]Cov[m,m]}},$$
$$Cov[f,m] = \frac{1}{|\Omega|}\sum_{i\in\Omega} f(i)m(i) - \frac{1}{|\Omega|^2}\sum_{i\in\Omega} f(i)\sum_{j\in\Omega} m(j) \tag{4}$$

$$L_{ncc} = 1 - \sum_{P\in\Omega} \frac{\left(\sum_{P_i}\left(f(P_i) - \hat{f(P)}\right)\left(m(P_i) - \hat{m(P)}\right)\right)^2}{\left(\sum_{P_i}\left(f(P_i) - \hat{f(P)}\right)^2\right)\left(\sum_{P_i}\left(m(P_i) - \hat{m(P)}\right)^2\right)} \tag{5}$$

Here, $f(P)$ and $\hat{f(P)}$ are the local patches in image f and the local mean of $f(P)$, respectively.

L_{reg} is the total variation loss [27] and is employed to prevent unrealistic deformation fields and overfitting. L_{reg} is calculated as follows:

$$L_{reg} = \frac{1}{3\Omega}\sum_{i\in\Omega}\sum_{n=1,2,3} |u(i+e_n) - u(i)| \tag{6}$$

where $u \in \Omega$ is the predicted deformation field and e_n is the standard basis of Ω.

Several stacked registration network models use a single L_{cc} as the

similarity loss, which is beneficial for network training and convergence. However, for the proposed method, L_{ncc} can provide a higher sensitivity for the local differences between the registration images. To ensure training efficiency and provide higher registration accuracy at the same time, L_{cc} was used in conjunction with $_{Lncc}$ in our method.

4. Experiments

4.1. State-of-the-art methods and experimental details

To demonstrate the effectiveness of the proposed method, extensive experiments were conducted on the registration of the large-scope abdominal CT images and chest CT images acquired at different respiratory phases, and the MRI brain-atlas-based registration. We compared the proposed method with a number of state-of-the-art deep-learning registration methods, including Voxelmorph [27], VTN [30], CCV [37], ViT-V-Net [41], TransMorph [45], and XMorpher [46], as well as two popular traditional registration methods, Elastix [13] and Ants [49].

The proposed method was implemented using TensorFlow 1.15 and trained with two NVIDIA RTX 2080Ti and CUDA 10.0. The Adam optimizer was selected. The initial learning rate was 10^{-4} and the learning rate halved for every epoch after the 3rd epoch. There were five epochs in total, with each epoch consisting of 20,000 batches. The batch size is 2 pairs per batch.

4.2. Datasets and preprocessing

Abdominal Dataset: The training data for the registration of abdominal CT images were obtained from the MSD [50] datasets. This dataset contained data pertaining to a variety of patients and had different imaging ranges. Hence, we manually removed the data covering a very small body range (such as those containing partial information of abdominal organs) or an excessively large body range (such as those including neck or leg). We selected 500 abdominal CT images for training, and the preprocessing before training included resampling to 160 × 160 × 128, intensity normalization (Hounsfield unit (HU) range to [-500, 300]), and data enhancement (small elastic deformation for the moving images). Notably, the [-500, 300] HU range can provide a stronger soft-tissue contrast than the general HU range [-1000, 1000]. Finally, pairwise registration for the entire abdominal training data was conducted to train the different DL registration models.

Chest Dataset: The training data for the registration of chest CT images were obtained from the LIDC-IDRI [51] datasets. Five hundred chest images were artificially screened to ensure consistency of the body range and subsequently downsampled to a size of 160 × 160 × 128. The general HU range of [-1000, 1000] was selected as the HU value of the chest CT images.

Brain Dataset: The training data for the registration of brain MR images were obtained from three public datasets, i.e., ADNI (66 volumes) [52], ABIDE (1287 volumes) [53], and ADHD (949 volumes) [54]. The Skull of these MR images was removed by FreeSurfer [55] and then all data were downsampled to a size of 128 × 128 × 128. Intensities are bias-corrected by Ants [49] and linearly normalized to [0,1] by min–max normalization. Here, we trained this model by random pairwise registration, but not the atlas-based registration, because the pairwise registration is more general than the atlas-based registration [30].

4.3. Evaluation metrics

We evaluate the performance of the registration algorithms with four evaluation metrics, i.e., Dice score (Dice), Hausdorff Distance (HD), Average Symmetric Surface Distance (ASSD) and Mean Squared Error (MSE). They are defined below:

(1) Dice

The Dice score is based on the segmentation of some anatomical structures in the warped moving image and the reference image. The Dice score is computed as

$$Dice(V_f,\ V_m) = \frac{2 \times \left|V_f \cap V_m\right|}{\left|V_f\right| + \left|V_m\right|} \tag{7}$$

where V_f and V_m are volumes of the segmented anatomical structures in the fixed and warped moving images, respectively. Perfect registration with fully overlapped anatomical structures would obtain a Dice score of one, whereas a poor registration would result in a low Dice score.

(2) HD

The Hausdorff Distance measures the maximum distance between surfaces of the segmented anatomical structures in the warped moving image and the reference image, which is formulated as:

$$HD(S_f, S_m) = \max_{f \in S_f} \left\{ \min_{m \in S_m} \{ \left\| S_f - S_m \right\| \} \right\} \tag{8}$$

where S_f and S_m indicate all the voxels on the surfaces of segmented anatomical structures in the fixed and warped moving images, respectively.

(3) ASSD

Similar to HD, ASSD measures the average distance between surfaces of the segmented anatomical structures in the warped moving image and the reference image, which is formulated as:

$$ASSD(S_f, S_m) = \frac{1}{\left|S_f\right| + \left|S_m\right|} \left(\sum_{f \in S_f} \min_{m \in S_m} \|f - m\| + \sum_{m \in S_m} \min_{f \in S_f} \|m - f\| \right) \tag{9}$$

where S_f and S_m indicate all the voxels on the surfaces of segmented anatomical structures in the fixed and warped moving images, respectively.

(4) MSE

The MSE measures the average intensity difference between the warped moving image and the reference image, which is formulated as:

$$MSE\left(I_f, I_m\right) = \frac{1}{N} \sum \left| I_f - I_m \right|^2 \tag{10}$$

where I_f and I_m indicate intensity of all the voxels in the fixed and warped moving images, respectively. N is the image size.

5. Results

5.1. Comparison of abdominal CT registration performance

The testing data for abdominal CT registration were obtained from Tongji Hospital, affiliated with the Tongji Medical College of Huazhong University of Science & Technology. The data included 18 pairs of abdominal CT images and are also preprocessed as same as the training data. A pair of images represented two images collected from one patient at two different times (the acquisition interval varied from several weeks to several months). All the test data were desensitized and contained the segmentation information of the liver (LIV), portal vein (POV), hepatic vein (HEV), aorta (AOR), stomach (STO), gallbladder (GAL), heart (HEA), left kidney (LKID), spleen (SPL), and right kidney (RKID), which were labeled by an experienced radiologist as the ground truth for Dice score and HD computation. Table 2 and Table 3 list, respectively, the average registration Dice and HD for each labeled organ and all labeled organs. In each column, the number in bold indicates the best value.

Table 2
Dice results for the abdominal CT registration.

Methods	LIV	LKID	SPL	POV	HEV	AOR	RKID	STO	GAL	HEA	Mean
Ants	0.914	0.917	0.882	0.500	0.220	0.697	0.926	0.669	**0.478**	0.853	0.706
Elastix	0.923	0.922	**0.899**	0.514	0.207	**0.721**	**0.936**	0.673	0.432	0.864	0.709
Voxelmorph	0.894	0.901	0.855	0.418	0.153	0.674	0.911	0.632	0.419	0.889	0.675
VTN	0.919	0.915	0.880	0.449	0.167	0.691	0.924	0.655	0.427	0.915	0.694
CCV	0.891	0.91	0.867	0.498	0.204	0.694	0.918	0.600	0.453	0.843	0.688
ViT-V-Net	0.907	0.903	0.862	0.459	0.164	0.674	0.906	0.652	0.431	0.862	0.682
TransMorph	0.917	**0.923**	0.884	0.486	0.194	0.699	0.935	0.656	0.440	0.902	0.704
XMorpher	0.918	0.915	0.876	0.493	0.196	0.703	0.922	0.662	0.448	0.894	0.703
SAN	**0.924**	0.918	0.882	**0.567**	**0.277**	0.719	0.930	**0.675**	0.439	**0.919**	**0.725**

Table 3
HD results for the abdominal CT registration.

Methods	LIV	LKID	SPL	POV	HEV	AOR	RKID	STO	GAL	HEA	Mean
Ants	10.21	7.12	12.63	**13.73**	17.85	**15.64**	13.87	11.17	7.65	8.28	11.82
Elastix	10.54	7.45	13.28	14.12	17.96	15.90	12.58	11.56	7.44	**7.86**	11.87
Voxelmorph	10.46	7.93	12.03	15.67	19.88	20.64	12.89	12.96	8.88	9.78	13.11
VTN	10.08	6.98	11.58	14.30	18.03	19.43	12.14	11.83	**3.79**	8.40	11.66
CCV	9.93	7.36	11.60	14.85	19.10	19.74	11.70	12.58	5.77	9.51	12.21
ViT-V-Net	10.18	7.25	11.56	14.68	18.65	19.79	11.94	12.61	6.38	9.59	12.26
TransMorph	10.04	6.59	**11.30**	14.43	18.58	19.23	10.33	12.16	4.82	9.29	11.78
XMorpher	9.76	6.85	11.73	14.79	18.21	19.66	11.92	12.43	5.34	9.57	12.03
SAN	**9.23**	**6.39**	11.56	13.75	**17.37**	15.71	**9.83**	**11.12**	4.18	8.09	**10.72**

As is evident from Table 2, the proposed method achieved the highest registration Dice score for the liver, portal vein, hepatic vein, aorta, stomach, and heart. For other organs,the left kidney, spleen, right kidney, and gallbladder, the proposed method provided competitive registration results compared to the other methods. The average Dice score of all organs using the proposed method was considerably higher than that using the other compared methods. The HD results in Table 3 shows the maximum surface error of organ registration. It is apparent that the proposed method provides the minimum HD on half of all organs. Also, the proposed method provides the minimum average HD for all organs.

Fig. 5 shows the registration results of the seven compared deep-learning registration methods (i.e., without Ants and Elastix) on the abdominal CT images. It depicts the coronal, sagittal and axial slices and

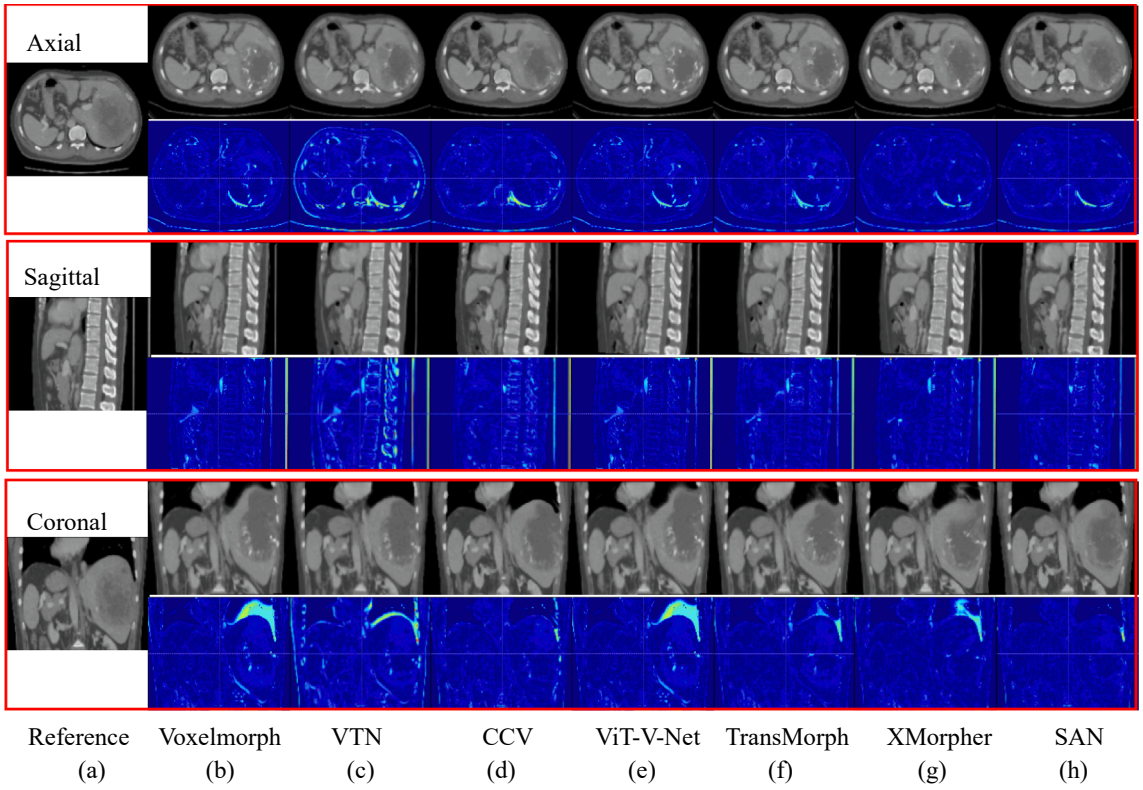

Fig. 5. Registration results for the axial, sagittal and coronal slices and the corresponding subtraction images on the abdominal CT data using Voxelmorph, VTN, CCV, ViT-V-Net, TransMorph, XMorpher, and SAN methods. (a) Reference images; (b)–(h): Warped moving images and subtraction images for Voxelmorph, VTN, CCV, ViT-V-Net, TransMorph, XMorpher and SAN methods, respectively.

the corresponding subtraction images that are obtained by calculating the mean absolute error between the reference image and the warped moving image for the different methods. The red and blue colors in the subtraction image means large and small differences of voxel intensity, respectively. Evidently, our method provides the most clear subtraction images by recovering the deformation of all the tissues effectively, which indicates the advantage of the proposed method in the registration of abdominal CT images.

5.2. Comparison of chest CT registration performance

Eight pairs of chest CT images for registration testing were also obtained from the Tongji hospital mentioned earlier. In the chest CT images, the heart (HEA), pulmonary artery (PA), aorta (AOR), left lung (LL), right lung (RL), trachea (TRA), and pulmonary vein (PV) were masked as segmentation ground truths by an experienced radiologist. The Dice and HD results of the testing data are listed in Table 4 and Table 5, respectively. The numbers in bold indicate the best values for the marked organs. As is evident, the average Dice and HD results obtained by the proposed method is better than those obtained by the other methods, especially for organs of the pulmonary artery, aorta, and pulmonary vein.

A visual comparison of the registration results is illustrated in Fig. 6. Evidently, all the compared methods can attenuate the initial misalignment of lung to some extent. However, as shown in Fig. 6, the SAN method has the best registration performance by providing the most clear subtraction image and the most similar registered result to the reference image among the seven methods.

5.3. Comparison of brain MRI registration performance

In addition to the abdominal and chest image registration, the MRI brain-atlas-based registration is also widely-used in clinical analysis. Thus we also evaluated our method for brain MR images atlas-based registration. The LPBA [56] dataset was used for brain MRI registration test. These data included 40 cases, and each case contained 56 segmentation labels as the ground truth for the evaluation. All the testing data are preprocessed as same as the training data and the first scan in LPBA is fixed as the atlas in our experiments. Table 6 shows the average values of Dice, HD, ASSD and MSE for all the compared methods. The proposed SAN method provides the highest Dice, ASSD and MSE value than other evaluated methods.

5.4. Ablation experiments

To evaluate the impact of the similarity attention model, we trained the model with and without it and conducted a simple ablation experiment using the abdominal CT images. The results are presented in Table 7, where the w-SA and wo-SA mean the proposed method with and without similarity attention model, respectively.

As is evident, using similarity attention module provided better registration results, especially for portal vein, hepatic vein, aorta, gallbladder and heart. This indicates that the similarity attention module can improve the sensitivity of the network to image differences and improve registration accuracy.

6. Discussions

Numerical Comparison and Discussion. Tables 2-6 provide quantitative numerical comparison results, in terms of the Dice, HD, ASSD and MSE, between the state-of-the-art methods and our SAN method. Compared with two traditional iteration-based methods, the proposed method provides slight improvement in accuracy and great improvement in execution time.

Compared with six DL-based methods, the proposed method provides more accurate registration results. Particularly, from Table 2 of the abdominal CT dataset, we can observe that our method improves the average Dice by 5 % compared to VoxelMorph, and by 2.1–4.3 % compared to other state-of-the-art methods. This performance gain is also consistent on the chest CT and brain MRI dataset; the average Dice of the proposed SAN model is 0.7–2.8 % higher on the chest CT dataset, and 0.9–3.2 % higher on the brain MRI dataset than other state-of-the-art methods.

Meanwhile, from the results in Table 2 and Table 4, we can see that for some difficult organs such as portal vein, hepatic vein, aorta, stomach, pulmonary artery and pulmonary vein, our method outperforms all other state-of-the-art methods by a larger margin. This indicates that our method has clear advantages over other state-of-the-art methods for the organs that are difficult to align. The reasons for this result are twofold. On one hand, the multi-scale image similarity in our method is non-local and based on the learned features. This strategy is more sensitive and accurate for the match/mismatch between images due to the learned high-dimensional inherent features. On the other hand, the displacement searching space is used as the soft attention to assist the network training and prediction, it can thus reduce the difficulty of deformation field prediction and improve the registration accuracy and robustness for the difficult task.

Visual Comparison and Discussion. From Figs. 5 and 6, we can see that the registered images obtained by our method have the highest image-wise similarity than those obtained by other compared methods. This implies that most regions with the same anatomic semantics are well-aligned by our registration method. Other methods fail to match fine details in several regions, especially those with large deformation.

Summary. Overall, our method has several advantages. Firstly, we observed that our method achieves considerable performance gain over the other state-of-the-art registration methods for various medical image registration tasks. The multi-scale similarity attention modular in our method effectively captures the image similarity to build a displacement searching space for the assistance of displacement prediction and, therefore, it makes the registration model more sensitive and robust to the disalignment in the images to be registered.

Secondly, the attention or transformer-based registration methods in the literature usually use the attention mechanism to learn and extract image features. Such strategy requires a large amount of data to ensure good performance. Unlike those methods, our similarity attention modular is to process the extracted features and thus can reduce to some

Table 4
Dice results for the chest CT registration.

Methods	HEA	PA	AOR	LL	RL	TRA	PV	Mean
Ants	**0.951**	0.603	0.873	0.933	0.946	0.777	0.498	0.797
Elastix	0.951	0.608	0.877	0.936	0.948	**0.778**	0.506	0.801
Voxelmorph	0.938	0.582	0.848	0.942	0.945	0.746	0.503	0.786
VTN	0.947	0.610	0.851	**0.963**	**0.964**	0.771	0.500	0.801
CCV	0.941	0.568	0.864	0.949	0.953	0.709	0.488	0.782
ViT-V-Net	0.936	0.593	0.844	0.935	0.939	0.752	0.488	0.784
TransMorph	0.949	0.607	0.879	0.954	0.955	0.767	0.507	0.803
XMorpher	0.948	0.611	0.862	0.958	0.960	0.773	0.511	0.803
SAN	0.950	**0.621**	**0.888**	0.954	0.956	0.772	**0.532**	**0.810**

Table 5
HD results for the chest CT registration.

Methods	HEA	PA	AOR	LL	RL	TRA	PV	Mean
Ants	**4.17**	15.30	7.16	12.61	9.66	**14.07**	16.45	11.35
Elastix	4.54	15.12	7.84	12.77	9.74	14.57	16.63	11.60
Voxelmorph	4.50	14.57	8.90	11.82	10.28	15.59	16.85	11.79
VTN	4.73	14.07	8.00	9.61	9.49	14.87	15.90	10.95
CV	6.83	15.21	9.80	9.54	9.36	15.64	15.76	11.73
ViT-V-Net	5.66	14.41	9.47	10.46	10.48	15.69	16.77	11.85
TransMorph	5.13	14.09	7.92	10.01	9.86	14.87	15.68	11.08
XMorpher	5.24	14.20	8.13	9.93	9.64	14.85	15.90	11.13
SAN	4.51	**13.75**	**7.54**	**9.37**	**8.23**	14.26	**15.65**	**10.47**

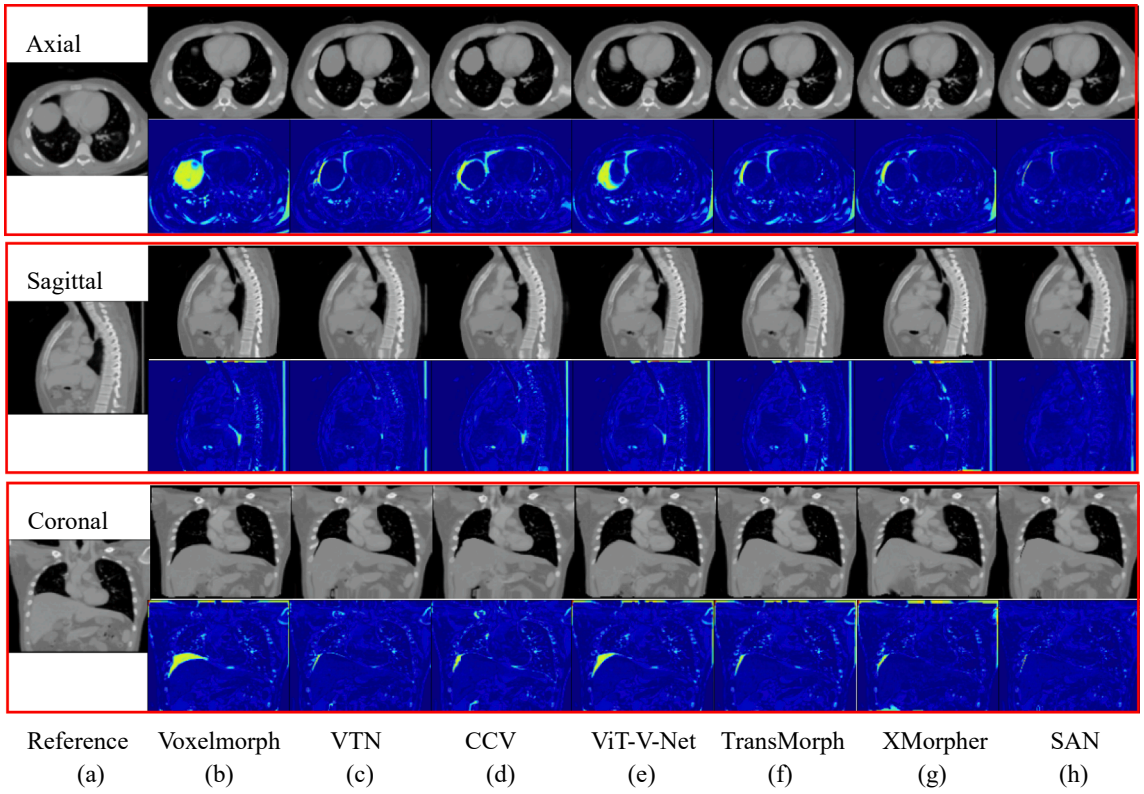

Fig. 6. Registration results for the axial, sagittal and coronal slices and the corresponding subtraction images on the chest CT data for Voxelmorph, VTN, CCV, ViT-V-Net, TransMorph, XMorpher and SAN methods. (a) Reference images; (b)–(h): Warped moving image and subtraction image for Voxelmorph, VTN, CCV, ViT-V-Net, TransMorph, XMorpher and SAN methods, respectively.

Table 6
Registration results for the brain MRI data.

Methods	Dice	HD	ASSD	MSE
Ants	0.708 ± 0.015	**4.885 ± 0.359**	1.138 ± 0.082	6.441 ± 0.894
Elastix	0.674 ± 0.013	4.953 ± 0.326	1.258 ± 0.069	9.564 ± 0.659
Voxelmorph	0.688 ± 0.015	4.997 ± 0.362	1.208 ± 0.077	6.947 ± 0.578
VTN	0.703 ± 0.014	4.969 ± 0.333	1.286 ± 0.076	6.869 ± 0.489
CV	0.696 ± 0.015	4.986 ± 0.355	1.335 ± 0.077	6.994 ± 0.688
ViT-V-Net	0.680 ± 0.015	5.013 ± 0.350	1.272 ± 0.076	9.135 ± 0.747
TransMorph	0.695 ± 0.014	4.936 ± 0.336	1.267 ± 0.075	6.793 ± 0.696
XMorpher	0.696 ± 0.013	4.974 ± 0.338	1.226 ± 0.074	6.772 ± 0.625
SAN	**0.712 ± 0.012**	4.934 ± 0.322	**1.094 ± 0.068**	**5.963 ± 0.511**

extent the impact of data size.

Thirdly, a coarse-to-fine global registration strategy in our method addresses the challenge of large search range for the non-local image similarity and calculation cost. Therefore, it is robust to the large deformation. Most of the methods in the literature do not have the coarse-to-fine strategy and may not be robust to the large deformation.

7. Conclusion

In this study, we proposed a similarity-based local attention model and embedded it into a multi-scale CNN for 3D medical image registration. The designed similarity-based local attention model can determine the similarity between the reference and moving images and guide the prediction of the deformation field. In combination with the multi-scale structure, similarity-based local attention becomes non-local. The multi-scale CNN can also provide a more lightweight network and coarse-to-fine registration. We conducted extensive experiments, including three application scenarios and comparison with eight state-of-the-art registration methods. The experimental results demonstrated

Table 7
The registration results with and without similarity attention on abdominal CT images.

Methods	LIV	LKID	SPL	POV	HEV	AOR	RKID	STO	GAL	HEA	Mean
wo-SA	0.910	0.914	0.880	0.502	0.207	0.701	0.927	0.663	0.412	0.868	0.698
w-SA	**0.924**	**0.918**	**0.882**	**0.567**	**0.277**	**0.719**	**0.930**	**0.675**	**0.439**	**0.919**	**0.725**

that the proposed method provides better performance in terms of registration accuracy and robustness compared with the other compared methods.

CRediT authorship contribution statement

Fei Zhu: Conceptualization, Data curation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Sheng Wang:** Data curation, Validation. **Dun Li:** Supervision, Investigation, Writing – review & editing. **Qiang Li:** Supervision, Conceptualization, Methodology, Resources, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

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

Data availability

For the open source datasets, I have shared data links in the references of the article, for the private datasets, I have no permission to share.

References

[1] T. De Silva, A. Uneri, M.D. Ketcha, S. Reaungamornrat, G. Kleinszig, S. Vogt, N. Aygun, S.F. Lo, J.P. Wolinsky, J.H. Siewerdsen, 3D–2D image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch, Phys. Med. Biol. 61 (8) (2016) 3009–3025.

[2] D. Sarrut, Deformable registration for image-guided radiation therapy, Z. Med. Phys. 16 (4) (2006) 285–297.

[3] R. Chandrashekara, A. Rao, G.I. Sanchez-Ortiz, R.H. Mohiaddin, D. Rueckert, Construction of a statistical model for cardiac motion analysis using nonrigid image registration, in: C. Taylor, J.A. Noble (Eds.), information Processing in Medical Imaging, Springer, Berlin, Heidelberg, 2003, pp. 599–610.

[4] Y.B. Fu, C.K. Chui, C.L. Teo, E. Kobayashi, Motion tracking and strain map computation for quasi-static magnetic resonance elastography, in: G. Fichtinger, A. Martel, T. Peters (Eds.), Medical Image Computing and Computer-Assisted Intervention, Springer, Berlin, Heidelberg, 2011, pp. 428–435.

[5] J.E. Iglesias, M.R. Sabuncu, Multi-atlas segmentation of biomedical images: A survey, Med. Image Anal. 24 (1) (2015) 205–219.

[6] X. Yang, B. Fei, 3D prostate segmentation of ultrasound images combining longitudinal image registration and machine learning, In: Proc. SPIE 8316: Medical Imaging 2012, San Diego, California, United States, 2012, 83162O.

[7] E.S. Andersen, K. Noe, T.S. Srensen, S.K. Nielsen, L. Fokdal, M. Paludan, J. C. Lindegaard, K. Tanderup, Simple DVH parameter addition as compared to deformable registration for bladder dose accumulation in cervix cancer brachytherapy, Radiother. Oncol. 107 (1) (2013) 52–57.

[8] M. Velec, J.L. Moseley, C.L. Eccles, T. Craig, M.B. Sharpe, L.A. Dawson, K.K. Brock, Effect of breathing motion on radiotherapy dose accumulation in the abdomen using deformable registration, Int. J. Radiat. Oncol.*Biol.*Phys. 80 (1) (2011) 265–272.

[9] F. Zhu, M. Ding, X. Zhang, Self-similarity inspired local descriptor for non-rigid multi-modal image registration, Inf. Sci. 372 (1) (2016) 16–31.

[10] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: a survey, IEEE Trans. Med. Imaging 32 (1) (2013) 1153–1190.

[11] T. Vercauteren, X. Pennec, A. Perchant, N. Ayache, Diffeomorphic demons: Efficient non-parametric image registration, NeuroImage 45 (1) (2009) 61–72.

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

[13] S. Klein, M. Staring, K. Murphy, M.A. Viergever, J.P. Pluim, Elastix: A toolbox for intensity based medical image registration, IEEE Trans. Med. Imaging 29 (1) (2010) 196–205.

[14] M. Gong, S. Zhao, L. Jiao, D. Tian, S. Wang, A novel coarse-to-fine scheme for automatic image registration based on sift and mutual information, IEEE Trans. Geosci. Remote Sens. 52 (7) (2014) 4328–4338.

[15] R. Szeliski, J. Coughlan, Spline-based image registration, Int. J. Comput. Vision 22 (3) (1997) 199–218.

[16] M. Simonovsky, B. Gutirrez-Becker, D. Mateus, N. Navab, N. Komodakis, A deep metric for multimodal registration, in: S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, W. Wells (Eds.), In: Medical Image Computing and Computer Assisted Intervention, Springer, Cham, 2016, pp. 10–18.

[17] G. Haskins, J. Kruecker, U. Kruger, S. Xu, P.A. Pinto, B.J. Wood, P.K. Yan, Learning deep similarity metric for 3D MR-TRUS image registration, Int. J. Comput. Assist. Radiol. Surg. 14 (3) (2019) 417–425.

[18] M.P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F.V. Gleeson, S.M. Brady, J. A. Schnabel, MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration, Med. Image Anal. 16 (7) (2012) 1423–1435.

[19] Y. Sun, A. Moelker, W.J. Niessen, T.V. Walsum, Towards robust ctultrasound registration using deep learning methods, in: D. Stoyanov (Ed.), Understanding and interpreting Machine Learning in Medical Image Computing Applications, Springer, Cham, 2018, pp. 43–51.

[20] K. A. J. Eppenhof, M. W. Lafarge, P. Moeskops, M. Veta, J. P. W. Pluim, Deformable image registration using convolutional neural networks, In: Proc. SPIE 10574, Medical Imaging 2018: Image Processing, Houston, Texas, United States, 2018, 105740S.

[21] T. Sentker, F. Madesta, R. Werner, GDL-FIRE^{4D}: Deep learning-based fast 4D CT image registration, in: A. Frangi, J. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention, Springer, Cham, 2018, pp. 765–773.

[22] M. Modat, G.R. Ridgway, Z.A. Taylor, M. Lehmann, J. Barnes, D.J. Hawkes, N. C. Fox, S. Ourselin, Fast free-form deformation using graphics processing units, Comput. Methods Programs Biomed. 98 (3) (2010) 278–284.

[23] J.A. Shackleford, N. Kandasamy, G.C. Sharp, On developing B-spline registration algorithms for multi-core processors, Phys. Med. Biol. 55 (21) (2010) 6329–6351.

[24] R. Werner, A. Schmidt-Richberg, H. Handels, J. Ehrhardt, Estimation of lung motion fields in 4D CT data by variational non-linear intensity-based registration: A comparison and evaluation study, Phys. Med. Biol. 59 (15) (2014) 4247–4260.

[25] R. Castillo, E. Castillo, R. Guerra, V.E. Johnson, T. McPhail, A.K. Garg, T. Guerrero, A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets, Phys. Med. Biol. 54 (7) (2009) 1849–1870.

[26] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, A. V. Dalca, An unsupervised learning model for deformable medical image registration, In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 9252-9260.

[27] G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A.V. Dalca, Voxelmorph: A learning framework for deformable medical image registration, IEEE Trans. Med. Imaging 38 (8) (2019) 1788–1800.

[28] J. Zhang, Inverse-consistent deep networks for unsupervised deformable image registration, ArXiv: 1809.03443 (2018).

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

[30] S. Zhao, T.F. Lau, J. Luo, E. Chang, Y. Xu, Unsupervised 3d end-to-end medical image registration with volume tweening network, IEEE J. Biomed. Health. Inf. 24 (5) (2019) 1394–1404.

[31] S. Zhao, Y. Dong, E. Chang, Y. Xu, Recursive Cascaded Networks for Unsupervised Medical Image Registration, In: 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019, pp. 10599-10609.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, In: Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[33] D. Sun, X. Yang, M. Y. Liu, J. Kautz, PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume, In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 8934-8943.

[34] T.C.W. Mok, A.C.S. Chung, Large Deformation Diffeomorphic Image Registration with Laplacian Pyramid Networks, in: A.L. Martel (Ed.), Medical Image Computing and Computer Assisted Intervention, Springer, Cham, 2020, pp. 211–221.

[35] X. Hu, M. Kang, W. Huang, M.R. Scott, R. Wiest, M. Reyes, Dual-Stream Pyramid Registration Network, in: D. Shen (Ed.), Medical Image Computing and Computer Assisted Intervention, Springer, Cham, 2019, pp. 382–390.

[36] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, M. Gelautz, Fast cost-volume filtering for visual correspondence and beyond, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 504–511.

[37] X. He, J. Guo, X. Zhang, H. Bi, S. Gerard, D. Kaczka, A. Motahari, E. Hoffman, J. Reinhardt, R.G. Barr, E. Angelini, A. Laine, Recursive Refinement Network for Deformable Lung Registration between Exhale and Inhale CT Scans, ArXiv: 2106.07608 (2021).

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv:2010.11929 (2020).

[39] Z. Zu, G. Zhang, Y. Peng, et al., VAN: Voting and Attention Based Network for Unsupervised Medical Image Registration, in: Pacific Rim International Conference on Artificial Intelligence, Springer, Cham, 2021, pp. 382–393.

[40] J.Q. Zheng, Z. Wang, B. Huang, et al., Recursive Deformable Image Registration Network with Mutual Attention, in: Annual Conference on Medical Image Understanding and Analysis, Springer, Cham, 2022, pp. 75–86.
[41] J. Chen, Y. He, E.C. Frey, et al., Vit-v-net: Vision transformer for unsupervised volumetric medical image registration, ArXiv: 2104.06468 (2021).
[42] L. Liu, Z. Huang, P. Liò, et al., Pc-swinmorph: Patch representation for unsupervised medical image registration and segmentation, ArXiv:2203.05684 (2022).
[43] T. C. W. Mok, A. Chung, Affine Medical Image Registration with Coarse-to-Fine Vision Transformer, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20835-20844.
[44] Y. Lei, Y. Fu, T. Wang, et al., 4D-CT deformable image registration using multiscale unsupervised deep learning, Phys. Med. Biol. 65 (8) (2020), 085003.
[45] J. Chen, E.C. Frey, Y. He, et al., TransMorph: Transformer for unsupervised medical image registration, Med. Image Anal. 82 (2022), 102615.
[46] J. Shi, Y. He, Y. Kong, et al., XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2022, pp. 217–226.
[47] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: 2013 ICML International Conference on Machine Learning, Atlanta, USA, 2013, p. 3.
[48] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, ArXiv: 1506.02025 (2015).
[49] B.B. Avants, N. Tustison, G. Song, Advanced normalization tools (Ants), Insight J 2 (365) (2009) 1–35.
[50] MSD, Medical segmentation decathlon, <https://medicaldecathlon.com/>.
[51] LIDC-IDRI, The Lung Image Database Consortium Image Collection, <https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI>.
[52] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C.R. Jack, W. Jagust, J. Q. Trojanowski, A.W. Toga, L. Beckett, Ways toward an early diagnosis in alzheimers disease: the alzheimers disease neuroimaging initiative (ADNI), Alzheimer's Dementia 1 (1) (2005) 55–66.
[53] A.D. Martino, C. Yan, Q. Li, E. Denio, F.X. Castellanos, K. Alaerts, J.S. Anderson, M. Assaf, S.Y. Bookheimer, M. Dapretto, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, Mol. Psychiatry 19 (6) (2014) 659–667.
[54] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D.S. Margulies, R. C. Craddock, The neuro bureau ADHD-200 preprocessed repository, Neuroimage 144 (B) (2017) 275–286.
[55] B. Fischl, Freesurfer, Neuroimage 62 (2) (2012) 774–781.
[56] D.W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K.L. Narr, R. A. Poldrack, R.M. Bilder, A.W. Toga, Construction of a 3D probabilistic atlas of human cortical structures, Neuroimage 39 (3) (2008) 1064–1080.