Multi-domain abdomen image alignment based on multi-scale diffeomorphic jointed network*

LU Zhengwei, WANG Yong, GUAN Qiu, CHEN Yizhou, LIU Dongchun, and XU Xinli**

College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310014, China

(Received 31 March 2022; Revised 22 June 2022)

Recently, the generative adversarial network (GAN) has been extensively applied to the cross-modality conversion of medical images and has shown outstanding performance than other image conversion algorithms. Hence, we propose a novel GAN-based multi-domain registration method named multiscale diffeomorphic jointed network of registration and synthesis (MDJRS-Net). The deviation of the generator of the GAN-based approach affects the alignment phase, so a joint training strategy is introduced to improve the performance of the generator, which feedbacks the structural loss contained in the deformation field. Meanwhile, the nature of diffeomorphism can enable the network to generate deformation fields with more anatomical properties. The average dice score (*Dice*) is improved by 1.95% for the computer tomography venous (CTV) to magnetic resonance imaging (MRI) registration task and by 1.92% for the CTV to computer tomography plain (CTP) task compared with the other methods.

Document code: A **Article ID:** 1673-1905(2022)10-0628-7

DOI https://doi.org/10.1007/s11801-022-2052-3

Many different domains of images can provide different information, especially in the field of medical imaging. The multi-domain medical image of the abdomen often contains multi-modality and multi-phase images. By aligning the spatial structure of the multi-domain images, the doctor can use this information to make better diagnoses and treatment plans for their patients[1]. Due to the internal or external forces, the anatomical structures of the abdomen in the multi-domain images are inevitably deformed, so the acquired multi-domain images are usually misaligned[2]. If the register can precisely calculate the spatial correspondence between two images, it can align them effectively[3,4]. However, in the cross-domain image of abdomen, due to significant style differences, accurate spatial deformation field generation requires auxiliary information of labels[5-7]. Meanwhile, labeling is always time-consuming and costly, which makes it very difficult to use labels to assist in registration[8,9].

Multi-domain images can be generated without paired images with generative adversarial network (GAN), providing a new solution to multi-domain alignment tasks that lack label data. In this paper, we propose a joint training strategy. First, we pre-train a registration network on the real image set. Based on the property that the pre-trained registration network can generate reasonable deformation fields between the real image pairs while generating unreasonable deformation fields on the synthetic and real image pairs with structural deviations, we introduce the adversarial learning mechanism after the register to feed the structural deviations between the real image and the synthesized image caused by the synthesizer during the registration process, thus guiding the synthesizer to generate the structure of the synthesized images more consistent. In addition, due to the large deformations in the abdominal images, we construct a kind of multi-scale diffeomorphic registration network to enhance the performance of multi-domain registration in the abdomen and generate deformation fields with a reasonable anatomical structure much better.

The diffeomorphic maps are used to preserve the topology structure. In some literature, the deformation can be represented as a member of the Lie algebra so that all the deformation field is defined as φ^t, where t is set from 0 to 1 and is exponentiated to produce a time one deformation φ^1. In this paper, the φ^1 is obtained by integrating the stationary velocity field v over time $t \in [0, 1]$ using the scaling and squaring method for the image pairs. Following the Padé approximant, as shown in Eq.(1), in general, $T>6$, and in our experiment, T is set as 7. To obtain the φ^1, we recursively calculate Eq.(2)

$$\varphi^{(1/2^T)} = x + v(x)/2^T, \tag{1}$$

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

Traditional deformable registration is always ab iterative-based procedure that gradually increases the similarity between images to complete the registration task. All these methods minimize a function that measures the morphological differences between the pairs. Some of

* This work has been supported by the National Natural Science Foundation of China (Nos.U20A20171, 61802347, 61972347, and 61773348), and the Science Foundation of Zhejiang Province (Nos.LY21F020027, LGF20H180002, and LSD19H180003).

** E-mail: gq@zjut.edu.cn

the studies parameterize the problem with displacement fields, such as Demons[10], B-splines-based methods' and their diffeomorphic implements. Besides, several studies see the registration task as a fluid problem, including vector momentum-parameterized stationary velocity field[11] and large displacement diffeomorphic metric mapping[11]. These classic deformable registration methods solve registration problems by iterating optimization objectives. So, when the image pairs have great morphological differences, the registration time would increase dramatically.

Learning-based cross-domain registration methods need to build a similarity loss to measure the quality of a deformation field[12]. The similarity measure that is insensitive to pixel intensity difference of image is generally selected. However, when the stylistic differences in the image pairs are too large, using some of the mainstream metrics (cross-correlation coefficient, mutual information) does not yield satisfactory results.

ZHU et al[13] presented an approach for learning to translate an image from a source domain to a target domain in the absence of paired examples. GAN allows for stylistic migration of images while keeping their content as unchanged as possible. By converting the images to be registered to the same domain, the cross-domain registration is simplified to mono-domain registration.

Some researches used a generator network that was trained by unaligned images[13-17]. The generator should be able to transform a real image from one domain to the other, and the anatomical structures of the synthesized image and the original one should be aligned. Then, the register can be trained on the real/synthetic data in one domain. However, the lack of sufficient structural constraints usually results in large synthetic misalignments. To generate a more anatomically correct deformation field, registration loss is used as a constraint on the synthesizer in Ref.[18]. However, they found the registration performance was worse than conventional registration networks[19], and ARAR et al[20] also achieved success in the field of natural image registration by feeding the registration loss back into generator network. However, medical images are typically more structurally complex than natural images. Because the generator does not have explicit structural constraints in the adversarial learning of unpaired images, and although most of the structural information can be retained in the adversarial learning, subtle unreasonable structures may still be generated to have an impact on the subsequent mono-domain registration.

In the traditional registration methods, multi-scale and cascaded registration methods have achieved good results, so some works are trying to introduce these into learning-based registration methods. ZHAO et al[21] proposed recursive cascaded networks, a general architecture that enabled learning deep cascades for deformable image registration. Their experiments demonstrated that cascaded networks could achieve better performance than deeper or wider networks with the same number of parameters. KIM et al[22] proposed CycleMorph (CM) imposing the cycle consistency on images to improve topological preservation. And the multi-scale implementation aligned the images at patches and original ones. However, many experiments are needed to find the optimal patch size and overlapping. What's more, the training on patches of the volume is not easy to converge for large deformation organs.

The architecture of our proposed multiscale diffeomorphic jointed network of registration and synthesis (MDJRS-Net) is illustrated in Fig.1. In this research, we have two domains, X and Y. Let x and y be the samples that are randomly sampled from X and Y. We utilize GAN to translate the image from one domain to another. G is the generator. D and D_φ are the discriminators. R is the registration network. The real images x_i and y_j are from one patient but haven't been aligned. y_i^* is translated by the G from x_i.

The goal of the research is to align the image x_i to y_j and x_i and y_j are the paired but not aligned images.

The register R generates the deformation field φ between real image y_i and fake image y_j^* or real images y_i and y_j, shown as

$$\varphi_{i\to j} = \mathrm{R}(y_i, y_j), \quad \varphi_{i\to j^*} = \mathrm{R}(y_i, y_j^*), \tag{3}$$

where $\mathrm{R}(\cdot)$ denotes a multivariate function that generates a deformation field from a pair of prior distribution y_i, y_j. The transform function $\varphi_{i\to j}$ denotes the registration direction from moving image y_i to fixed image y_j.

The G is a generative network that generates a cross-domain image. The D_φ is a deformation field discriminator which improves the G using the strategy of adversarial learning from Ref.[23], shown as

$$L_{\mathrm{dadv}} = \left|\log(\mathrm{D}_\varphi(\varphi_{j\to i})) + \log(1-\mathrm{D}_\varphi(\varphi_{j^*\to i}))\right|. \tag{4}$$

G could be optimized by the difference of the two deformation fields that contain deviation of the synthesizer, shown as

$$L_{\mathrm{dc}} = \left\|\varphi_{i\to j}(y_i) - \varphi_{i\to j^*}(y_i)\right\|_1. \tag{5}$$

We utilize CycleGAN[13] to generate the cross-domain images, because it can perform on unaligned datasets. In the joint training, L_{adav} and L_{dc} are integrated into the training of G, shown as

$$L_{\mathrm{syn}} = L_{\mathrm{cyclegan}} + \lambda L_{\mathrm{dadv}} + \beta L_{\mathrm{dc}}. \tag{6}$$

As shown in Fig.2, our registration network consists of two different subnetworks based on VoxelMorph (VM)[19]. The g_0 and g_1 are parameter networks (structure of U-Net) to generate velocity fields. The low-resolution registration network has half number of parameters compared to the high-resolution one. The scaling and squaring layers are the same in the two subnetworks. L_{reg} consists of similarity loss and regular items. The loss function is shown as

$$L_{reg}=\lambda_1 L_{smooth}+\lambda_2 L_{sim}+\lambda_3 L_{anchor}, \quad (7)$$

where L_{smooth} is the smooth item to preserve the topology, and L_{anchor} is the edge loss to limit the deformation field to nearly a zero displacement at the edge of the images.

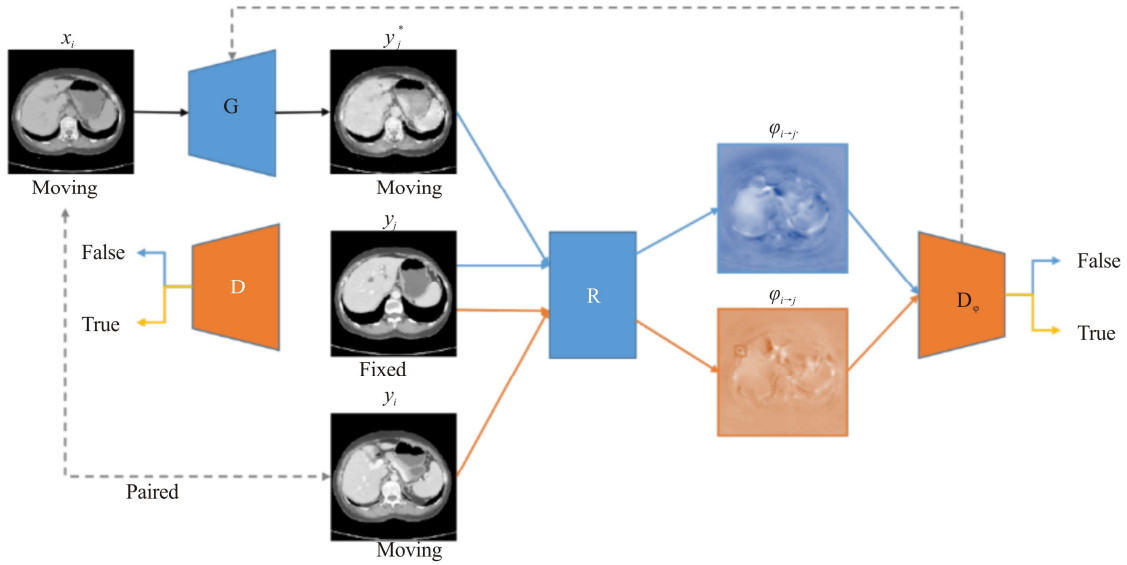

Fig.1 Overview of our MDJRS-Net (The stream of blue lines represents the registration of composite image and real image, and the orange ones are for real pairs)

The similarity loss L_{sim} is an indirect assessment to measure the quality of the transform field. L_{ncc} is always used when the images are cross-domain pairs, shown as

$$L_{ncc}=-\frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}}. \quad (8)$$

The Cov(·) denotes the covariance between image x and y, and the Var(·) denotes the variance of x.

In the neural network training stage, registration may crash because of the large displacement of the whole moving image, especially at the end of the training. To reduce this occurrence, anchor loss is proposed. The loss is shown as

$$L_{anchor}=\sum_{i}^{\Omega} x(i), \quad (9)$$

where the Ω denotes the pixel index set at the edge of the deformation field.

Like deep supervision, we use a cascaded supervision where the registration loss is applied for both low-resolution and high-resolution networks to accelerate the training process and balance the registration performance of the low-resolution network and high-resolution one.

The training stage refers to Ref.[24]. First, the registration network R is pre-trained on a real image dataset. Subsequently, we trained the G and R jointly and the training process is shown in Fig.1. Because of the low-quality synthetic images which would bring in a lot of noise interfering with R training, the Frechet inception distance (FID)[25] is applied to measure the quality of synthetic image.

To evaluate the registration results, the average dice coefficient (*Dice*), average symmetric surface distance (*ASD*), and the negative Jacobian determinant map ($|J_\varphi| \leqslant 0$) are used as evaluation metrics in this paper.

Dice measures the organ label's volume coverage rate after registration. Given the segment label of the moving image as A, and the segment label of the moved image as B, *Dice* can be defined as

$$Dice(A,B)=2\cdot\frac{A\cap B}{|A|+|B|}. \quad (10)$$

ASD also measures the organ label's coverage but calculates outline distances of the segment labels, shown as

$$ASD(A,B)=\frac{\sum_{a\in A}\min_{b\in B}\mathrm{d}(a,b)}{|A|}, \quad (11)$$

where d(·) is the Euclidean distance between the two pixels a and b.

The negative Jacobian determinant map is used to measure the organ folding caused by the deformation field φ. The smaller the value, the less the organ folding. The Jacobian determinant is shown as

$$\mathrm{jet}(\nabla\varphi)=\begin{vmatrix}\frac{\partial\varphi_u}{\partial x} & \frac{\partial\varphi_u}{\partial y}\\ \frac{\partial\varphi_v}{\partial x} & \frac{\partial\varphi_v}{\partial y}\end{vmatrix}, \quad (12)$$

where jet(·) calculates the determinant for the Jacobian of φ. φ_u and φ_v are the components of deformation φ.

Our abdominal dataset was obtained from the partner hospital and included computer tomography plain (CTP)/ computer tomography venous (CTV)/magnetic resonance

venous (MRV) three-dimensional (3D) voxel data from a total of 51 patients. We used a total of 1 020 pairs of two-dimensional (2D) slices of these 3D voxels as training data. To evaluate the performance of the alignment, we manually annotated the contours of the liver and spleen for all test samples.

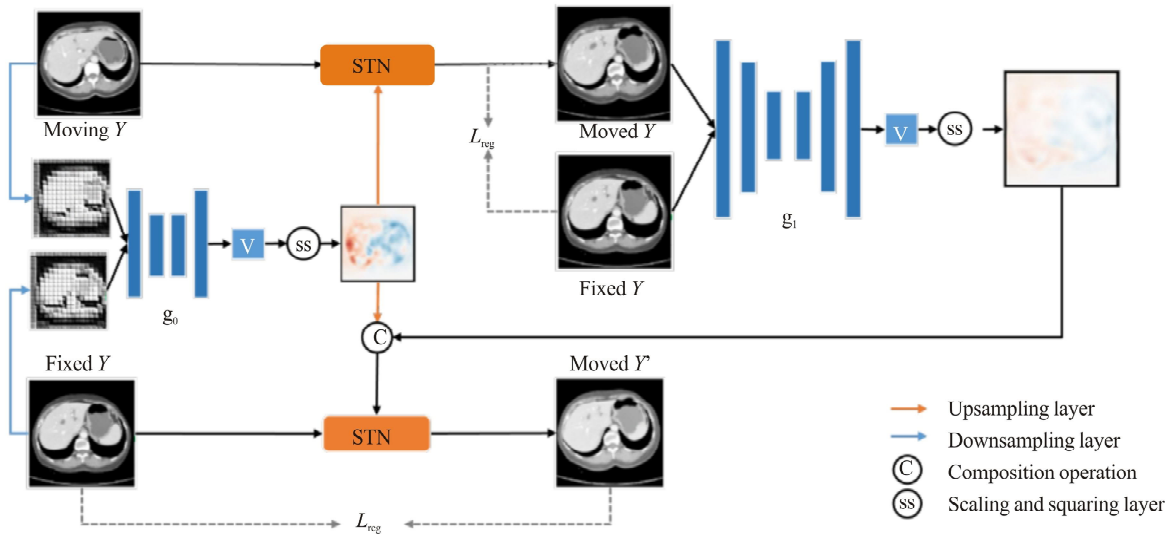

Fig.2 Registration network of MDJRS-Net (The spatial transform network[4] warps the moving image to the moved image, and the parameter networks g_0 and g_1 are learning to generate the deformation field)

Our experiments were conducted on two 2080Ti GPUs with 11GB of video memory. In our experiments, we set the batch size to 2 and trained 500 epochs iteratively during the training of the registration network and the joint network, respectively. We normalized these abdominal images using Z-Score before feeding into the network. The structure of the generator module G and the deformation field discriminator D_φ of the network is referred to Ref.[16], using the residual blocks of ResNet. For the registration module, we used U-Net as the backbone to generate the deformation fields. In the loss function of the registration network, the hyperparameter of the smoothing term is set to 3, the optimizer uses Adam, and the learning rate is set to 2×10^{-5}. The structure of the evaluation network is similar to that of InceptionV3 when the FID is then computed[25].

We compare the MDJRS-Net with five exiting multi-domain registration methods, including synthesis, localization, inpainting, and registration (SLIR)[16], unsupervised multi-modal deformable image registration (UMDIR)[15], adversarial uni- and multi-modal stream networks (ADSN)[17], geometry preserving registration network (GPN)[20], and CM[22], where CM does not translate the images between two domains. Due to the lack of segment masks are in our training dataset, the weak supervision methods[1,6,7,18] that require labels have not been compared. Besides, different structures of the registration network are applied, or the cascade supervision is ablated, and we test the performance of the MDJRS-Net on the CTP/CTV dataset.

The qualitative comparison of multi-modality and multi-phase registration is illuminated in Fig.3. The bi-direction alignments of the CTV/MRV image and CTP/CTV image are both carried out, and the organ contours coincidence of the moved image obtained by MDJRS-Net and the fixed image are much better than that of other methods. In addition, the moved images obtained by other comparison algorithms have more distortions during multi-modality alignment, which implies that the multi-modality alignment is of more challenging.

In Tab.1, we can find that although the network architecture of CM is based on GAN, its performance on the alignment of MRI and CT with large modal differences is much lower than that of other GAN-based cross-domain registration methods because it does not perform image style transformation on cross-domain image pairs.

The ablation study of multi-phase alignment is demonstrated as Fig.4. From these figures, when multi-scale layer or velocity integration in the MDJRS-Net model is ablated, the organ contour distance between the fixed image and the moved image increases, and the moved image may have distortion even.

In our ablation experiment, the hyperparameter of the smooth item in the registration loss is set to 3, and the hyperparameters of similarity loss and the anchor loss are set to 7 and 10.

The Jacobian determinant maps are shown in Fig.5, The first and the second columns show the moving image and the fixed image. The third column is the moved image warped by different registration networks. The fifth column is the $|J_\varphi|$ map, which shows the Jacobian determinants map of the deformation field. The red

points represent the organ folding which is anatomic.

The first row shows the multi-scale network trying to align in more detail, for example, in the lower right corner of the image, it looks better than the other two. However, the transform field causes too much organ folding (red points).

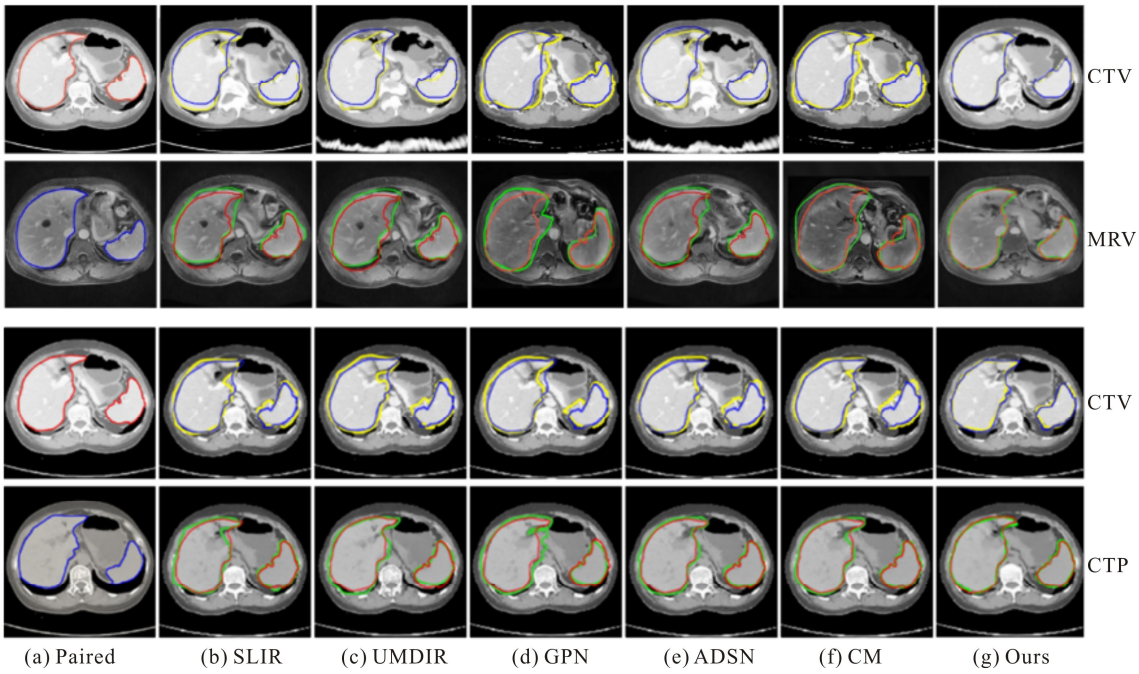

Fig.3 The results of comparison experiments for multi-modality and multi-phase registration (The four images of multi-modality and multi-phase in column (a) are paired respectively, and the red and blue lines in the images are the outlines of the liver and the spleen; Images in the columns (b—g) are the moved images, and the yellow and green lines in the images are the outlines of the organs that are warped by different registers)

The models in the second and the third rows are using the static velocity field to generate a smooth displacement field. Besides, because the upsampling (bilinear interpolation) is required for the low-resolution deformation field, it was smoothed implicitly. This operation also reduces the number of singularities in the generated velocity field.

Quantitative results are statistically in Tab.2. We can see that the multi-scale registration network improves the performance, and the constraint of the diffeomorphism would keep the topological property.

It shows that the multi-scale networks significantly improved performance but increased the folding of the organization. The folding organ abnormalities were reduced with the diffeomorphism constraint, while the registration results have deteriorated slightly.

Finally, we improve the performance of registration by adding cascade supervision and striking a balance between the rationality of the deformation field and the accuracy of the registration task.

In this research, a novel registration model named MDJRS-Net is proposed. First, the loss term is divided into two parts, and the part caused by synthesis deviation is only used to optimize the synthesizer. Secondly, the adaptive generalization loss is constructed to generalize

Tab.1 Average dice scores and average symmetric surface distances of multi-domain registration (The symbol "→" represents the transform direction from different domains)

Model	CTV→MRV	MRV→CTV
SLIR[16]	83.98±3.94	82.67±4.70
UMDIR[15]	83.82±3.99	82.14±5.12
GPN[20]	80.49±7.71	79.66±8.78
ADSN[17]	83.44±4.16	81.91±5.23
CM[22]	78.86±10.0	79.42±8.85
Ours	85.93±2.31	85.77±2.48
Model	CTV→CTP	CTP→CTV
SLIR[16]	89.71±2.31	90.33±2.22
UMDIR[15]	89.24±2.40	89.90±2.29
GPN[20]	89.26±2.40	89.45±2.37
ADSN[17]	90.49±2.13	91.12±1.95
CM[22]	90.13±2.26	91.03±1.99
Ours	92.41±1.31	92.75±2.08

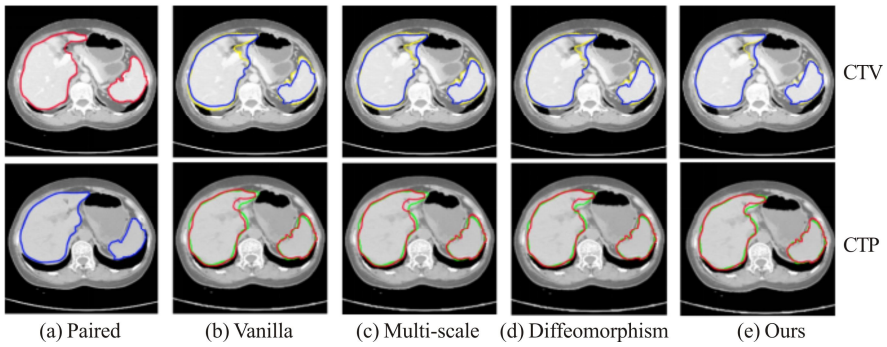

Fig.4 Outlines of the organs in ablation study of multi-phase registration for CTV/CTP

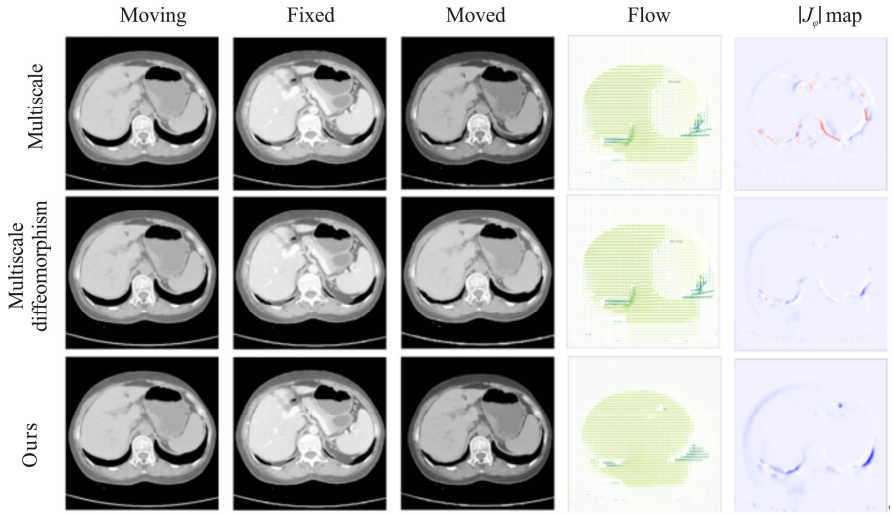

Fig.5 The deformation field visual results of multi-phase alignment ablation

the register. Thirdly, with our MDJRS-Net, the deformation can obtain better topological properties, and the performance gets boost to a certain extent. Finally, the results of experiments show that the performance of the proposed method in this paper can reach state-of-the-art.

Tab.2 Average dice scores and average number of voxels with non-positive Jacobian determinant of the ablation study of multi-phase registration by different registers

Model	CTV→CTP		CTP→CTV	
	Dice	$\lvert J_\varphi\rvert\leq 0$	*Dice*	$\lvert J_\varphi\rvert\leq 0$
VoxelMorph	91.62	70.23	92.08	73.03
Multi-scale	**92.52**	78.42	92.72	79.94
Multi-scale diffeomorphism	92.31	13.63	92.16	15.23
Multi-scale diffeomorphism cascade supervision	92.41	**8.72**	**92.75**	**7.52**

In this research, the quality evaluation threshold relies on experience which is difficult to generalize to other data sets. In future work, we will use the deep network to search the threshold at the training stage to adapt to any other data distribution.

Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

References

[1] LIU F, CAI J, HUO Y, et al. JSSR：a joint synthesis, segmentation, and registration system for 3D multi-modal image alignment of large-scale pathological CT scans[C]//European Conference on Computer Vision 2020, August 23-28, 2020, Glasgow, UK. Berlin, Heidelberg：Springer-Verlag, 2020：257-274.

[2] HUIJSKENS S C, VAN DIJK I W E M, VISSER J, et al. Abdominal organ position variation in children during image-guided radiotherapy[J]. Radiation oncology, 2018, 13(1)：1-9.

[3] ZHANG Y, JIANG F, SHEN R. Region-based face alignment with convolution neural network cascade[C]// 24th International Conference on Neural Information Processing, November 14-18, 2017, Guangzhou, China. Berlin, Heidelberg：Springer-Verlag, 2017：300-309.

[4] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28：2017-2025.

[5] CAO X, YANG J, GAO Y, et al. Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis[J]. Medical image analysis, 2017, 41：18-31.

[6] FAN J, CAO X, WANG Q, et al. Adversarial learning for mono- or multi-domain registration[J]. Medical image analysis, 2019, 58：101545.

[7] DUBOST F, DE BRUIJNE M, NARDIN M, et al. Multi-atlas image registration of clinical data with automated quality assessment using ventricle segmentation[J]. Medical image analysis, 2020, 63：101698.

[8] BLENDOWSKI M, HANSEN L, HEINRICH M P. Weakly-supervised learning of multi-modal features for regularised iterative descent in 3D image registration[J]. Medical image analysis, 2021, 67：101822.

[9] TANG T, GUAN Q, WU Y. Support vector machine incremental learning triggered by wrongly predicted samples[J]. Optoelectronics letters, 2018, 14(3)：232-235.

[10] VERCAUTEREN T, PENNEC X, PERCHANT A, et al. Diffeomorphic demons：efficient non-parametric image registration[J]. Neuroimage, 2009, 45(1)：S61-S72.

[11] BEG M F, MILLER M I, TROUVÉ A, et al. Computing large deformation metric mappings via geodesic flows of diffeomorphisms[J]. International journal of computer vision, 2005, 61(2)：139-157.

[12] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27：2672-2680.

[13] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York：IEEE, 2017：2223-2232.

[14] LIU M Y, BREUEL T, KAUTZ J. Unsupervised image-to-image translation networks[J]. Advances in neural information processing systems, 2017, 30：700-708.

[15] QIN C, SHI B, LIAO R, et al. Unsupervised deformable registration for multi-modal images via disentangled representations[C]//International Conference on Information Processing in Medical Imaging, June 2, 2019, Hong Kong, China. Berlin, Heidelberg：Springer-Verlag, 2019：249-261.

[16] WEI D, AHMAD S, HUO J, et al. SLIR：synthesis, localization, inpainting, and registration for image-guided thermal ablation of liver tumors[J]. Medical image analysis, 2020, 65：101763.

[17] XU Z, LUO J, YAN J, et al. Adversarial uni-and multi-modal stream networks for multimodal image registration[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention, October 4-8, 2020, Lima, Peru. Berlin, Heidelberg：Springer-Verlag, 2020：222-232.

[18] ZHOU B, AUGENFELD Z, CHAPIRO J, et al. Anatomy-guided multimodal registration by learning segmentation without ground truth：application to intraprocedural CBCT/MR liver segmentation and registration[J]. Medical image analysis, 2021, 71：102041.

[19] BALAKRISHNAN G, ZHAO A, SABUNCU M R, et al. Voxelmorph：a learning framework for deformable medical image registration[J]. IEEE transactions on medical imaging, 2019, 38(8)：1788-1800.

[20] ARAR M, GINGER Y, DANON D, et al. Unsupervised multi-modal image registration via geometry preserving image-to-image translation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-18, 2020, Seattle, WA, USA. New York：IEEE, 2020：13410-13419.

[21] ZHAO S, DONG Y, CHANG E I, et al. Recursive cascaded networks for unsupervised medical image registration[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea (South). New York：IEEE, 2019：10600-10610.

[22] KIM B, KIM D H, PARK S H, et al. Cyclemorph：cycle consistent unsupervised deformable image registration[J]. Medical image analysis, 2021, 71：102036.

[23] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30：6626-6637.

[24] CHEN Y, LU Z, YANG X H, et al. Multi-domain abdomen image alignment based on joint network of registration and synthesis[C]//International Conference on Neural Information Processing, December 8-12, 2021, Bali, Indonesia. Berlin, Heidelberg：Springer-Verlag, 2021：334-344.

[25] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 16-July 1, 2016, Las Vegas, USA. New York：IEEE, 2016：2818-2826.