A coarse-to-fine deformable transformation framework for unsupervised multi-contrast MR image registration with dual consistency constraint

Weijian Huang, Hao Yang, Xinfeng Liu, Cheng Li, Ian Zhang, Rongpin Wang, Hairong Zheng, Shanshan Wang, *Senior Member, IEEE*

***Abstract*— Multi-contrast magnetic resonance (MR) image registration is useful in the clinic to achieve fast and accurate imaging-based disease diagnosis and treatment planning. Nevertheless, the efficiency and performance of the existing registration algorithms can still be improved. In this paper, we propose a novel unsupervised learning-based framework to achieve accurate and efficient multi-contrast MR image registrations. Specifically, an end-to-end coarse-to-fine network architecture consisting of affine and deformable transformations is designed to improve the robustness and achieve end-to-end registration. Furthermore, a dual consistency constraint and a new prior knowledge-based loss function are developed to enhance the registration performances. The proposed method has been evaluated on a clinical dataset containing 555 cases, and encouraging performances have been achieved. Compared to the commonly utilized registration methods, including VoxelMorph, SyN, and LT-Net, the proposed method achieves better registration performance with a Dice score of 0.8397±0.0756 in identifying stroke lesions. With regards to the registration speed, our method is about 10 times faster than the most competitive method of SyN (Affine) when testing on a CPU. Moreover, we prove that our method can still perform well on more challenging tasks with lacking scanning information data, showing the high robustness for the clinical application.**

***Index Terms*— medical image analysis, multi-contrast, registration, unsupervised deep learning**

Manuscript received October 1, 2020; accepted February 09, 2021. This research was partly supported by Scientific and Technical Innovation 2030-"New Generation Artificial Intelligence" Project (2020AAA0104100, 2020AAA0104105), the National Natural Science Foundation of China (61871371, 81830056), Key-Area Research and Development Program of GuangDong Province (2018B010109009), the Basic Research Program of Shenzhen (JCYJ20180507182400762), Youth Innovation Promotion Association Program of Chinese Academy of Sciences (2019351).(Corresponding author: S. Wang)

W. Huang, H. Yang, L. Cheng, I. Zhang, H. Zheng, and S. Wang are with Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. S. Wang is also with Pengcheng Laboratory, Shenzhen, Guangdong, China and Pazhou Lab, Guangzhou, China (email: wj.huang@siat.ac.cn; hao.yang@siat.ac.cn; cheng.li6@siat.ac.cn; ianzhangfpv@gmail.com; hr.zheng@siat.ac.cn; sophiasswang@hotmail.com).

X. Liu and R. Wang are with Guizhou Provincial People's Hospital, Radiology Guiyang, Guizhou, China(email: lainiu6715613@163.com; wangrongpin@126.com).

I. Introduction

Multi-modal medical imaging plays an important role in many clinical applications [1]–[13]. Among them, multi-contrast magnetic resonance (MR) imaging is one of the most prevalent techniques as different MR imaging sequences can provide versatile information and highlight different regions of interest of the patient. For instance, diffusion-weighted imaging (DWI) and apparent diffusion coefficient (ADC) are functional MR images based on the movement of water molecules [14], [15]. T1-weighted images (T1), T2-weighted images (T2), and fluid-attenuated inversion-recovery (FLAIR) are structural MR images [16] which can indicate different characteristics of anatomical structures.

Multi-contrast MR imaging is of great significance for disease diagnosis and treatment response monitoring in clinical practices [17]–[20]. Structural MR images can clearly show the structures and boundaries of brain tissues but have moderate performances on discriminating brain lesions. On the other hand, functional MR images possess excellent capabilities of highlighting brain diseases, such as ischemic lesion regions. Analysis with multi-contrast images contributes to the comprehensive understanding of the patient. However, misalignment exists between different contrast images due to various issues of the scanning process, including physiological activities and eddy currents [21]. Fig. 1 shows the multi-contrast images of four examples. Physical space alignment has been conducted utilizing the provided scanning information. However, misalignment between the different contrast images can still be observed. In some cases, the scanning information might be lost due to data storage or transfer, large misalignment can happen. Misalignment brings difficulties to identify lesions accurately, which may have adverse effects on disease diagnosis. Multi-contrast MR image registration is needed.

Registration methods are available to alleviate the mismatch problem. Traditional multi-contrast registration algorithms rely on the interactive optimization process, which is not very applicable to the time-sensitive diagnosis required in clinical practices. Deep learning-based methods have been developed recently and can speed up the registration process at the cost

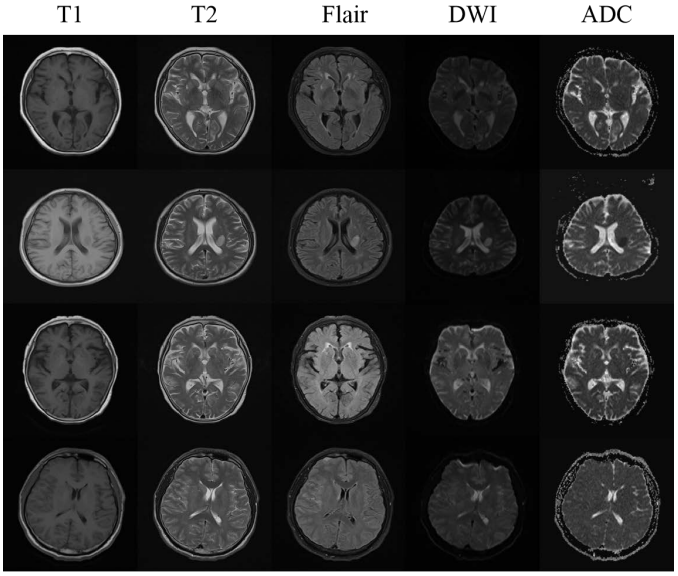

Fig. 1. MR multi-contrast brain images acquired from four candidates. There are differences between the contrasts that need to be registered.

of registration accuracy.

To achieve accurate and fast multi-contrast MR image registration, this paper proposes a novel concise registration framework. Specifically, we have made the following contributions:

1) We propose an unsupervised coarse-to-fine registration framework. A coarse registration is obtained by an affine transformation network, which is then refined by a subsequent deformable transformation network. These two transformations are integrated, and end-to-end image registration is achieved.
2) A dual consistency constraint is designed to maximize the cross-correlation of topology maps of multi-contrast MR images. The inverse deformation field is generated from the forward deformation field directly to reduce the time requirement. The designed consistency constraint is enforced on the bi-directional deformations so as to suppress pixel folding.
3) A prior knowledge-based loss function is designed to improve the sensitivity of mutual information (MI) for more accurate registration. Specifically, a negative area constraint is designed to limit signals that are registered in the fixed images background.
4) Extensive experiments with or without the first step of physical space alignment show the superiority of the proposed registration method compared to existing widely-employed approaches.

The rest of this paper is organized as follows: Section II introduces related work in medical image registration, Section III describes our methods, Section IV presents the experimental results and relevant analysis, and Section V gives the conclusion.

II. Related Work

A. Conventional image registration methods

Traditional image registration algorithms, such as elastic [22], [23], fluid [24]–[28] or B-spline models [29], are usually based on the iterative numerical solution of the optimization problem. Especially, in 1998, Thirion *et al.* [30] proposed a method called demons to estimate the velocity vector field between two adjacent images in a video. Specifically, they calculated the optical flow, used Gaussian filter to smooth the flow map, and optimized the predictions on each pair of images through multiple iterations. Since the successful implementation of demons, many variants were developed, such as the works by Wang *et al.* and Vercauteren *et al.* [31], [32]. In 2005, Beg *et al.* [24] proposed another famous registration algorithm, LDDMM (Large Displacement Diffeo-morphic Metric Mapping), by deducing and implementing the Euler-Lagrangian optimization to compute particle flows, solving a global variational problem, and estimating metrics for images. Subsequently, variants of this algorithm were also proposed, including Region-specific Diffeomorphic Metric Mapping (RDMM), vector momentum-parameterized Stationary Velocity Field (vSVF), and Symmetric image Normalization (SyN) [33]–[35]. Among them, SyN [35] has been the most widely employed algorithm in medical image registration.It described an Euler-Lagrange optimization based symmetric image normalization method for maximizing the cross-correlation. Nevertheless, the efficiency of these methods can still be improved since these methods are based on iterative optimization [4], [36].

B. Deep learning-based unimodal image registration

With the fast development in the deep learning field, some deep learning-based image registration models have been proposed. Initially, deep learning was employed to enhance the registration performance of the iterative methods. Then, deep reinforcement learning was introduced to predict steps of transformations until the optimal alignment was reached [37]–[40]. With the increased demand on the registration speed, deep learning-based registration methods were proposed [2], [41]–[43]. One representative work in this group is STN (Spatial Transform Network), which generates dense deformable transformations to register images. Since then, STN has been modified and utilized in various situations [44]. Yoo *et al.* [45] successfully employed STN to register electron microscopy images. They trained an autoencoder to reconstruct the fixed images and calculated a new loss between the reconstructed fixed images and the corresponding moving images. Krebs *et al.* [36], [46] proposed a random latent space learning method to alleviate the requirement on spatial regularization. De Vos *et al.* [41] developed a multi-stage and multi-scale approach to register unimodal images with a normalized cross correlation (NCC) loss and a bending energy regularization. However, this approach cascaded multiple networks, which severely increased the computational complexity. Balakrishnan *et al.* proposed the famous framework, VoxelMorph, and its derivative versions [2]–[5], which computed gradients of the transformation to backpropagate deformation errors during optimization. However, since the above methods all focus on

unimodal image registration, multi-contrast image registration remains to be explored.

C. Deep learning-based multi-modal image registration

Since multi-contrast MR image registration is similar to multi-modal medical image registration, we discuss multi-modal registration in this section to give a more comprehensive description. Compared with unimodal registration, multi-modal registration is more challenging because it is difficult to define effective similarity measures to guide local matching across different modalities. Mutual information (MI) is the most frequently utilized supervision in existing studies [47]. Li *et al.* [48] registered multi-modal retinal images by using the descriptor matching on the average phase map for global registration and using a deformable modality independent neighborhood descriptor method to locally optimize the registration results. Unfortunately, this method was based on manually designed features and it has limited robustness. Ceranka *et al.* [49] proposed a whole-body DWI and T1-weighted image registration method. This method roughly aligned the pelvis regions of the two modal images and then used MI to guide global registration. Cao *et al.* [50] developed an image synthesis-based method. They adopted a random forest to learn the transformation between computed tomography (CT) images and MR images, and synthesized pseudo CT images and pseudo MR images with similar anatomical structures. In this way, they transferred the multi-modal image registration task to a unimodal image registration task. Improved models over this original implementation were also proposed in [51]. Nonetheless, these methods require a robust domain transformation algorithm and their registration performances can be highly affected by the quality of the synthesized images [50].

III. Method

In this paper, we propose a concise registration algorithm for unsupervised multi-contrast MR image registration. The proposed method embeds an affine transformation network in a deformable network to achieve coarse-to-fine registrations. A dual consistency constraint is designed to further enhance the registration performance. Meanwhile, a prior knowledge-based guidance function is implemented. Here, let $K \in R$ represents the sample count in the multi-contrast datasets and $F \supset \left\{f^1, f^2 \cdots f^K\right\}$ and $M \supset \left\{m^1, m^2 \cdots m^K\right\}$ refer to the paired fixed image sets and moving image sets.

A. Affine transformation network – ATNet

STN [44] is a dynamic mechanism that can transform images or feature maps in a voxel-based manner. With this mechanism, a specific transformation can be performed all over the entire feature map, including scaling, cropping, rotating, etc. Owing to its high effectiveness, STN has been widely applied to deep learning-based registration tasks.

We use STN to perform affine transformation on the moving images [52], which geometrically consists of a non-singular linear transformation (transformation using a linear function). To clearly demonstrate the procedure, let $p\left(x_i, y_i\right)$ represent a pixel sampling from m, where x_i, y_i denotes as the coordinates of the corresponding pixel. Then the affine transformation can be expressed as:

$$A_\theta(p) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \cdot \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \tag{1}$$

where θ represents the parameters that determine the linear transformation. We pre-train a shallow regression network to predict those parameters. With the obtained parameters, STN can perform the affine transformation automatically without human involvement to roughly align the moving images M to corresponding fixed images F. This regress network is called the affine transformation network (ATNet) in our framework. With ATNet, we can acquire the affine transformed predictions of the original moving images, which are represented as $M_A \supset \left\{m_A^1, m_A^2 \cdots m_A^k\right\}$. These predictions are roughly aligned to F, and dense deformation transformations are needed to align the detailed local structures. It can be seen that only performing a linear transformation will not be able to capture the subtle differences between multi-contrast images. Besides, since affine transformations are global information-driven, the performance may be compromised when registered in low signals area. Therefore, predictions of the affine transformation network are treated as coarse registration images, which need to be further improved.

B. Deformable transformation network – DTNet

Deformable transformations are important for fine image registration. VoxelMorph [2]–[5] constructs a differentiable operation, which can be optimized through network training, on each pixel to realize image registration. Let us define φ as the obtained transformation field. Each value in φ represents an offset distance. Symbol $\circ$ refers to the transformation operator for m^k, which consists of pixel shifting and interpolation. For each pixel p in m^k transform to p' can be defined as:

$$p' = p + \varphi(p) \tag{2}$$

VoxelMorph performs an additional linear interpolation in neighboring pixels after the pixel transformation to avoid discontinuities in transformed images:

$$m \circ \varphi(p) = \sum_{q \in Z(p)} m(q) \prod_{dim \in x,y,z} \left(1 - \left|p_{dim} - q_{dim}\right|\right) \tag{3}$$

where Z represents the regions composed of adjacent pixels. Through this differentiable interpolation operation, the predicted results are smoother and more realistic.

We employ VoxelMorph as our deformable transformation network (DTNet) to conduct fine image registrations. Some changes in the network architecture were adopted. For example, we adopted a deeper convolution structure to fully extract features. In addition, the activate function of ReLu is replaced by Leaky ReLu. More details about the architecture will be arranged in section IV.

C. Coarse-to-fine multi-contrast image registration framework

To reduce the challenges of unsupervised multi-contrasts image transformation, we propose a coarse-to-fine image registration framework. Specifically, we embed the pre-trained ATNet $\mathrm{D}_\theta(F, M)$, with frozen parameters into DTNet. The affine transformed predictions M_A can serve as the inputs to DTNet. In this way, DTNet receives images that were roughly aligned to the fixed images with decreased image discrepancies. Different from existing methods that conduct two-step registrations of using affine transformations as preprocessing and then refine the predictions, the proposed framework adopts an end-to-end approach that conducts those operations in one architecture. Compared with the existing registration method, our method does not need to iterate over affine or deformation transformations. Meanwhile, we can obtain the affine transformed predictions and deformable transformed predictions as side outputs of the framework.

D. Dual consistency-constrained bi-directional image transformation

Intuitively, the registration procedure should be symmetrical, which refers to the bi-directional transformations between the moving images and the fixed images. This assumption was first proposed in [35] with an Euler Lagrange equation for iterative optimization and achieved great success in medical image registration. Inspired by this work, we propose a bi-directional image transformation method.

As defined in the previous section, φ is the transformation field for the forward transformation of registering moving images to fixed images. However, to inverse the transformation and restore the moving images, simply apply $-\varphi$ to the predictions of the forward deformation will not work because the correspondence between φ and the image pixels has been destroyed by the forward shift. Let $\varphi_{i,j}$ be the displacement of the pixel (i, j) in the moving image. After the forward transformation, pixel (i, j) becomes pixel $(i^{'}, j^{'})$ in the registered image. Then, $-\varphi_{i,j}$ should be the inverse displacement of the pixel $(i^{'}, j^{'})$ instead of (i, j), and we need to find the correspondence between $-\varphi_{i,j}$ and (i, j). Accordingly, we constructed the inverse deformation field φ^{-1} to guide the backward transformation. Instead of building a new network to generate φ^{-1} from scratch [53] or integrate the negative velocity field [5], we extrapolate φ^{-1} from φ to reduce complicated operations. Specifically, as φ consists of the horizontal and vertical offsets in the 2D space, we first decompose φ to obtain the two offset fields φ_x and φ_y, respectively. Then, we warp the offset fields with the original φ to form the deformed offset fields. By recombining the deformed offset fields, a new transformation field is generated. Finally, the inverse transformation field φ^{-1} is obtained by multiplying with -1. In this way, we successfully align the transformation field with the pixels in the registered images. To sum up, the whole process can be represented by the following equation:

$$\varphi^{-1} = -\Sigma_{x,y}\left(\varphi_i \circ \varphi\right) \tag{4}$$

Since there are no reference images to evaluate the accuracy of the multi-contrast registration predictions, it is difficult to conduct the bi-directional registrations simultaneously from M to F and from F to M. To combat this issue, we come up with a compromised solution that transfers the multi-contrast bi-directional image registration task to a unimodal image registration task, i.e. we use the predictions $M_D \supset \left\{m_d^1, m_d^2 \cdots m_d^k\right\}$ instead of the fixed images F to calculate the inverse transformed images: $M_D^{-1} = \mathrm{M}_D \circ \varphi^{-1}$. Here, we assume that M_D^{-1} should still maintain the same distribution as M_A. Base on this, we use a consistency loss to accurate constraint M_D^{-1} to M_A, which can be MSE or NCC. We can then obtain our integrated framework, the coarse-to-fine multi-contrast image registration framework with dual consistency constraint.

E. Coarse-to-fine multi-contrast image registration framework with dual consistency constraint

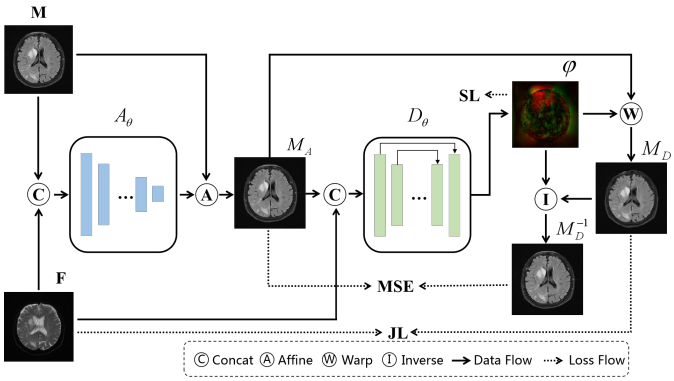

Fig. 2. The proposed coarse-to-fine multi-contrast image registration framework with dual consistency constraint. $\boldsymbol{A_\theta}$ represents the pre-trained ATNet. $\boldsymbol{D_\theta}$ refers to the DTNet. $\boldsymbol{M_D}$ and $\boldsymbol{M_D^{-1}}$ construct the bi-directional registration cycle.

Our coarse-to-fine multi-contrast image registration framework with dual consistency constraint is illustrated in Fig. 2. The framework consists of three main parts: 1) The pre-trained affine transformation network ATNet (A_θ) for coarse affine registration. The input to ATNet is a pair of M and F MR multi-contrast images. The output is the affine transformation for coarse alignment from M to F. The coarsely aligned images M_A are the inputs to the subsequent deformable transformation network. It is important to note that once the pre-training is finished, the parameters of ATNet are frozen and no longer updated. 2) The deformable transformation network DTNet is to generate the final predictions. The input to DTNet is a concatenation of F and M_A. The output is a densely transformation field φ. With φ, the final prediction M_D is generated. 3) A dual consistency constraint. We propose a novel inverse transformation from M_D to M_D^{-1} to further enhance the registration performance. We calculated the inverse transformation field φ^{-1} and warp M_D with it to obtain M_D^{-1}. By enforcing a similarity measure between M_D^{-1} and M_A, we achieve the dual consistency constraint. With the bi-directional registration strategy, undesirable interpolation during image

registration is expected to be suppressed and a more accurate registration can be obtained.

F. Loss function

As indicated in Fig. 2, multiple loss functions are utilized to optimize the multi-contrast MR image registration framework. For simplicity, we use $\xi_\theta(\cdot)$ to represent an undefined network which can be either ATNet (A_θ) or DTNet (D_θ).

The most important loss function used is Mutual Information (MI), which can measure the distribution dependence between two random variables [45]. Here, we define two marginal probability distributions, $p_F(f)$ and $p_M(m)$, and a joint probability distribution $p_{F,M}(f,m)$. MI measures the degree of dependence between F and M by calculating the distance between the joint distribution $p_{F,M}(f,m)$ and the distribution $p_F(f)p_M(m)$ by means of the Kullback-Leibler measurement [54]. MI loss (MI) can be written as Eq. 5:

$$MI(F,M) = -\iint p_{F,M}(f,m)\log(\frac{p_{F,M}(f,m)}{p_F(f)p_M(m)})dxdy \tag{5}$$

If F and M are independent, $p_{F,M}(f,m)$ is equal to $p_F(f)p_M(m)$, and $MI(F,M)$ will be zero, which means that there is no mutual information between the two variables. Maximization of MI is a general and powerful criterion because no assumptions are made regarding the nature of this dependence and no limiting constraints are imposed on the image content of different modalities involved [47].

Since MR images are usually in grayscale with background values close to 0, we suggest no signals should appear in the background regions of registered images. Based on this, we propose a prior knowledge-based background suppressing loss function: $MSE(f,m) = (f-m)^2$ when f are background pixels.

Combing the MI loss function and the prior knowledge-based background suppressing loss function, we obtain the first loss function, which is called a prior knowledge-based joint loss function ($\mathrm{JL}(F, \xi_\theta(F,M), \lambda)$) as shown in Eq.6:

$$\begin{aligned}\mathrm{JL}(F,\xi_\theta(F,&M),\alpha,\beta) \quad = \\ &\sum_{f,m}(\alpha MI(f,\xi_\theta(f,m))+ \\ &\beta\sum_i \begin{cases} MSE(f_i,\xi_\theta(f,m)_i), & \text{if} \quad f_i < \gamma \\ 0, & \text{otherwise} \end{cases}\end{aligned} \tag{6}$$

where $i \in N$ represents the pixels in images, γ is a threshold obtained from the data set to determine whether the pixel is background or not, α and β are adjust factors to balance the two losses. JL can not only constrain the global image alignment by maximizing MI, but also penalize the incorrect predictions in defined regions. This makes the predictions more in line with the nature of medical images.

The second loss function we use is to meet the dual consistency constraint. A simple MSE loss is calculated instead of MI loss between M_D^{-1} and M_A. The utilization of MSE loss is not fixed and can be replaced by similar losses, such as NCC or L_1-norm.

The last loss function is calculated to constrain the transformation field φ. Transformation may occur with an irregular displacement without constraint, whereas the above mentioned two losses can still be small through the interpolation algorithm. To prevent such situations, a spatially smooth loss function is designed to refine the transformation field φ:

$$\mathrm{SL}(\varphi) = \sum_{f,m} |\nabla\varphi(f,m)|^2 \tag{7}$$

where $\nabla(\cdot)$ represent the calculation of gradients. By limiting the gradient of the deformation field, we make sure that the transformation field is smooth, and extreme pixel displacement can be avoided.

The overall loss function to optimize the framework is calculated as shown in

$$\begin{aligned}\mathrm{Loss}_{total}(F,M) =&\lambda_1 SL(\varphi) + JL(F, D_\theta(F,M), \lambda_2, \lambda_3)+ \\ &\lambda_4 MSE(\mathrm{A}_\theta(F,M), \mathrm{D}_\theta^{-1}(F,M))\end{aligned} \tag{8}$$

The equation contains four adjust factors $\lambda_{i\in\{1,2,3,4\}}$. These are hyper-parameters that can be set to different values according to the experiment.

IV. Experiments and Results

In this section, we verify the effectiveness of the proposed methods through extensive experiments. In clinical practices, FLAIR and DWI are the most commonly used MR weighted sequences. Thus, our image registration experiments are mainly conducted with FLAIR and DWI data.

A. Dataset

The multi-contrast MR data were collected by Guizhou Provincial People's Hospital. This retrospective study was approved by the institutional review board of the hospital with the written informed consent requirement waived. All patient records were de-identified before analysis and reviewed by the institutional review boards to guarantee no potential risk to patients. The researchers who conduct the registration tasks have no link to the patients to prevent any possible breach of confidentiality.

In total, data from 555 patients are utilized with or without stroke lesions. Each patient was scanned with five sequences: T1 weighted, T2 weighted, FLAIR, ADC, and DWI. All images were obtained with a Siemens 1.5T scanner. Of the 555 cases, 466 are provided with the scanning information while the others are not. Two sets of experiments were conducted, one with the physical space alignment according to the provided scanning information and the other without. In the first set of experiments, only the 466 cases with scanning information were utilized. 426 cases were randomly selected as the training set and the remaining 40 cases as the test set. Stroke lesions in DWI and FLAIR images of the test set were annotated by experienced clinicians for quantitative result evaluation. In the second set of experiments, all 555 cases were used. Since no physical space pre-alignment was performed, the 89 cases without scanning information were included in the training set. All the data are resized to 224×224 with intensity normalized $[0,1]$.

B. Implementation details

Theoretically, ATNet and DTNet can adopt various network structures. In this study, we prefer simple network structures to reduce computational complexity. We will show in the results section that even with the selected simple network structures, our proposed method can still achieve good registration performance.

ATNet is implemented with a regression network, which contains five downsampling blocks and two fully-connected layers. Each downsampling block consists of two 3×3 convolutional layers followed by a 2×2 max pooling layer. The convolution operation is always followed by batch normalization and leaky ReLU activation unless otherwise specified. Finally, two fully-connected layers is appended to generate the 6 transformation parameters. With these parameters, affine transformations are performed. The channels of the downsampling blocks and the last two fully-connected layers are set as 16, 32, 32, 64, 64, 128, 32, and 6, respectively. ATNet has about 589k trainable parameters.

DTNet is modified from the famous UNet with an encoder-decoder architecture [55]. The encoder of DTNet is the same as the above mentioned ATNet, whereas the decoder is designed symmetrically to the encoder. For the last layer, we utilized two 3×3 convolutions with linear activations and then, the final transformation field φ can be obtained. DTNet has about 1478k trainable parameters.

Multiple comparison methods are adopted, including VoxelMorph (VM) [4], VoxelMorph-diff (VM-diff) [5], LT-Net [53], and Symmetric Normalization (SyN) [35]. VM is the most famous deep learning-based registration algorithms developed in recent years. We slightly adjust the method (using the MI loss) to make it suitable for multi-contrast image registration. For LT-Net, we discarded the label transfer part and only kept the main registration framework with the inverse module. SyN is a top-performing brain registration algorithm. It is implemented in the publicly available Advanced Normalization Tools (ANTs) software package [56] with a MI constraint for multi-contrast MR image registration. In our implementation, SyN has two designs: 1) Moving images go through ANTs-based affine transformations and SyN, represented as 'SyN(Affine)'; 2) Moving images go through SyN only, represented as 'SyN(Only)'. Since the GPU implementations for these two methods are not currently available, CPU implementations are utilized and the registration speed is reported accordingly.

Our method is implemented using Keras with a Tensorflow backend on a NVIDIA Titan Xp GPU. All experiments are based on 2D slices. During training, data augmentation methods are applied including random translations, rotations, dilations, and horizontal flip. The batch size is set to 32, and the learning rate is set to 0.01 with an Adam optimizer. Pre-training ATNet takes about 25 minutes, and the entire framework including DTNet requires another 20 minutes to optimize. The four weights in the loss function, $\lambda_{i\in\{1,2,3,4\}}$, were set to 1, 4, 100, 100 empirically. The threshold factors γ in the JL was set to 0.1. Our code will be available online at `https://github.com/SZUHvern/TMI_multi-contrast-registration`.

C. Results of multi-contrast MR image registration

In this section, qualitative and quantitative image registration results are reported. Quantitative results are calculated with regard to the alignment of stroke lesions between registered moving images and fixed images. Please note that there is still a lack of measurement metrics to characterize multi-contrast MR image registration. Although the area or shape of the stroke lesions may be differently presented in multi-contrast images, we believe that alignment between the stroke lesions can still reflect the registration performance.

We evaluated our method using Dice, Recall, and Precision, which are commonly used in computer vision. They are important indicators to assess the overall difference between our predictions and the ground truths. The exact formulas to calculate the three scores are: $Dice = 2TP/(2TP + FP + FN)$, $Recall = TP/(TP + FN)$, and $Precision = TP/(TP + FP)$. Here, TP (true positive) indicates the numbers of correctly register pixels, FP (false positive) indicates the numbers of pixels that the model register negative as positive, and FN (false negative) indicates the numbers of pixels that the model register positive as negative. We calculate these scores based for each case individually and report the average results. In addition, in order to quantify the deformation regularity, we calculate the Jacobian determinant J_φ as the derivative of the deformation field, and $|J_\varphi| < 0$ indicates the locations where folding has occurred. We report the number of pixels where $|J_\varphi| < 0$.

Example predictions of different methods are shown in Fig. 3. For example (a), the comparison methods generated deformed skulls while our method can keep the structure very well. In regions with low signals, such as example (b), iterative methods, SyN (Affine) and SyN (Only), show unexpected deformations and the tissues look abnormal. In example (c), the deep learning-based comparison methods cannot fully register the stroke lesions. Our method can still perform well thanks to the more robust registration flow we designed. Finally, when scanning artifacts exist in remote regions (example (d)), SyN (Only) shows obvious image distortion in order to fit the artifacts. Overall, satisfactory results are achieved by all the registration methods, and our method performs especially well for challenging cases with artifacts, sharp changes, etc.

The quantitative results are listed in Table I. Without registration, stroke lesion annotations in FLAIR and DWI images are misaligned with an average Dice score of 0.7822, which reflects the need for multi-contrast image registration. ATNet gets 0.8067, reflecting that even with the physical space alignment, linear transformations is still needed to achieve accurate registration. The methods without the affine transformation, SyN (Only) generate a similar result of 0.8101. It is improved to 0.8157 by introducing the affine transformation (SyN (Affine)). Our method achieves the highest score of 0.8397, proving its effectiveness in handling the multi-contrast image registration problem.

Efficiencies of the different methods are also compared. For fair comparisons, all the methods are tested on a CPU.

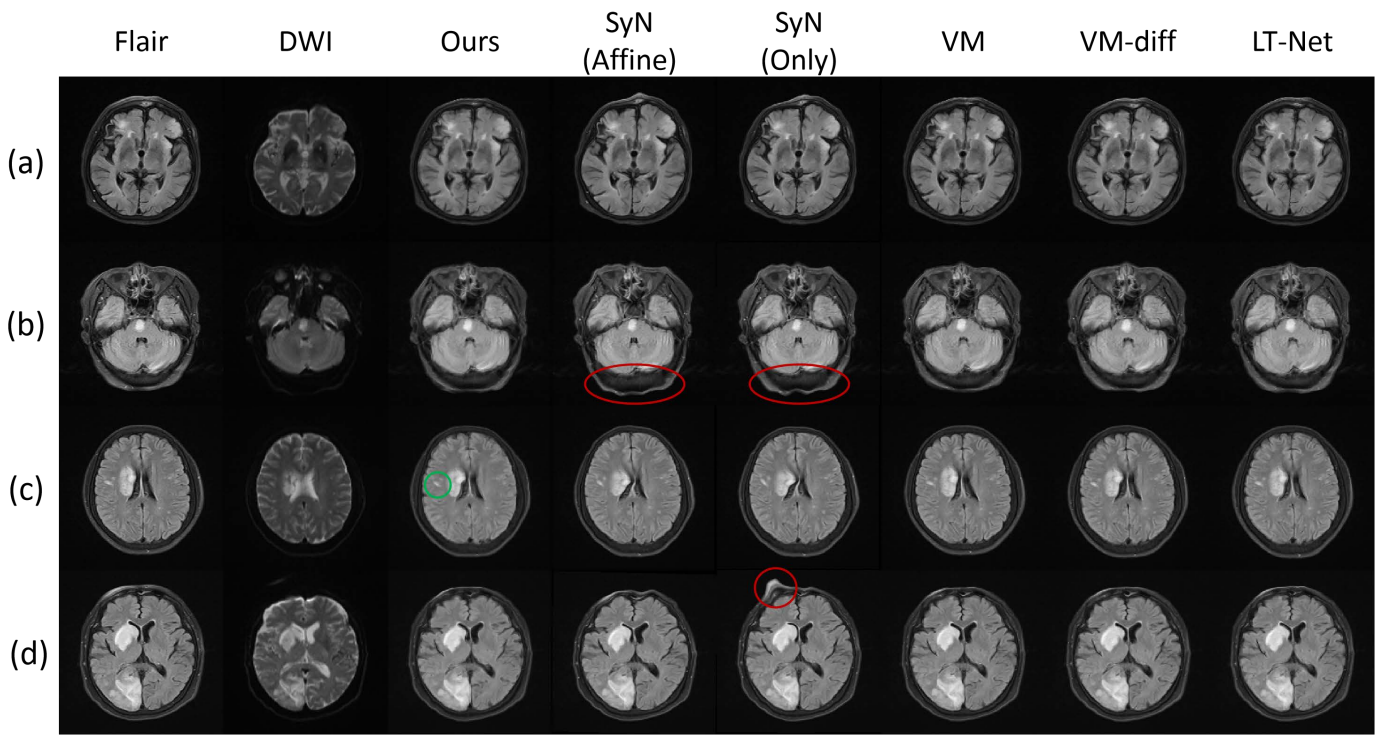

Fig. 3. Qualitative registration results of different methods. The green circle denotes regions that our method registered better than other methods while the red circle denotes unexpected predictions based on the specific comparison methods.

TABLE I
QUANTITATIVE MEASUREMENT OF THE STROKE REGION REGISTRATION RESULTS AND THE REQUIRED TEST TIME.

Method	Dice	Precision	Recall	Sec/Slice (GPU)	Sec/Slice (CPU)
Undef	0.7822 ± 0.0974	0.8491 ± 0.1007	0.7326 ± 0.1175	-	-
SyN(Only)	0.8101 ± 0.0979	0.8734 ± 0.1137	0.7669 ± 0.1237	-	1.4446
SyN(Affine)	0.8157 ± 0.0950	0.8769 ± 0.1061	0.7736 ± 0.1244	-	2.0335
LT-Net	0.7960 ± 0.1043	0.8480 ± 0.1087	0.7595 ± 0.1220	0.0162	0.2156
VM-diff	0.8011 ± 0.0991	0.8305 ± 0.1059	0.7826 ± 0.1184	0.0405	0.1801
VM	0.8053 ± 0.0954	$\mathbf{0.8861 \pm 0.0856}$	0.7534 ± 0.1322	0.0109	0.1699
ATNet	0.8067 ± 0.0935	0.8699 ± 0.0948	0.7590 ± 0.1128	**0.0100**	**0.0299**
Ours	$\mathbf{0.8397 \pm 0.0756}$	0.8856 ± 0.0808	$\mathbf{0.8081 \pm 0.1069}$	0.0223	0.2037

SyN (Affine) is the least efficient method that spends 2.0335 seconds to register one image slice, and ATNet has the highest efficiency which needs only 0.02 seconds. Comparing with the most competitive method, SyN (Affine), our method is about 10 times faster with better registration results. It can achieve the registration of one 3D image case (20 slices) within 5 seconds, which is sufficient for real-time diagnosis in clinical practices. The time spent can be further shortened to within 0.5 second/case when testing on a GPU.

D. Visualization of the transformation field

Visualizations of example transformation fields φ are shown in Fig. 4. These examples indicate that even after the physical alignment and affine transformation, large deformations (indicated by the red and green signals in the transformation fields) are still needed for accurate registrations. As a result, physical alignment, affine transformation, and deformable transformation, especially the latter two, are simultaneously required in applications.

E. Time consumption analysis of inverse transformation

To investigate the efficiency of the proposed inverse transformation, we compare the time consumption of our proposed method with the existing inverse methods, VM-diff [5] and LT-Net [53].

VM-diff [5] introduced an inverse deformation by adding a differential and integral layer (to generate velocity field) combined with a spatial transformation layer. The inverse deformation field is then obtained by iterating the negative velocity field. Specifically, this method split the registration into T=7 integration steps and then warp moving images according to the computed diffeomorphic field φ^{-1} using a spatial transform layer. Comparing with our method of calculating the inverse deformation field in one step, this method requires more operation steps.

LT-Net [53] is a cycle-correspondence learning method for atlas-based segmentation. This method builds a new network to learn the inverse deformable field and achieves the inverse transformation through a transformation layer. Since the in-

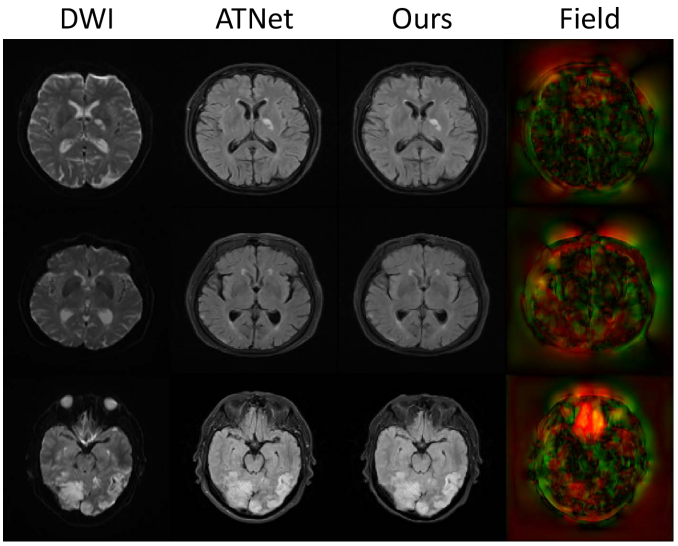

Fig. 4. Visualizations of the transformation fields φ. Red color indicates the transformation in the horizontal direction and green indicates the transformation in the vertical direction. Higher red or green color signals indicate larger transformations.

verse deformation field is realized by a new network, this method is more complicated than ours.

We conduct experiments to quantitatively compare the time consumption of different methods to prove our analysis (Table II). For fair comparisons, we implemented all methods with the same neural networks (DTNet) except for the inverse operation, and thus, the obtained time consumptions are different from those shown in Table I. As expected, our method is the fastest with a registration speed of 0.0233 s/slice on a GPU.

TABLE II
QUANTITATIVE TIME COMPARISON BETWEEN DIFFERENT INVERSE METHODS.

Method	Sec/Slice(GPU)	Sec/Slice(CPU)
VM-diff	0.0565	0.2147
LT-Net	0.0320	0.2479
Ours	**0.0223**	**0.2037**

F. Ablation experiment

We also conducted extensive ablation experiments to verify the effectiveness of the proposed framework. Firstly, we investigate the influence of network widths on the registration performance under two learning rates. Then, we inspect the importance of the all the proposed structures. Finally, we discussed the influence of JL's parameter selection on the prediction result.

In Fig. 5, we show the Dice scores of networks with different widths under two learning rates. Although the larger learning rate can lead to relatively faster convergence, fluctuated Dice score curves indicate that the training is unstable. Especially for width of 32 network, a smaller learning rate might be more appropriate. For the different network widths, significantly worse performance is observed with a width of 8 and 16, which might indicate that the network is not able to capture the complex image properties. Wider networks with widths of 32 and 64 show similar performance and the network with a width of 32 performs slightly better. It is worth noting that there is no overfitting in all implementations, which indirectly proves the suitability of our method for the multi-contrast image registration task.

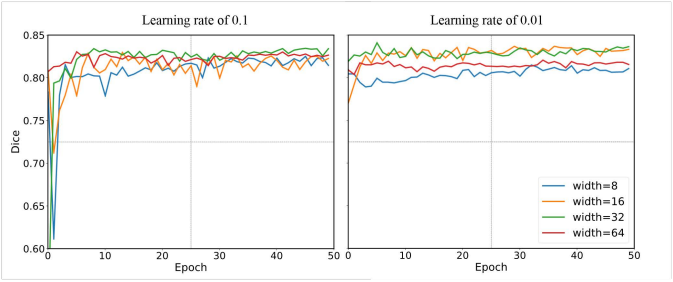

Fig. 5. Results of networks with different widths (8, 16, 32, 64) under two learning rates of 0.1 and 0.01. The width value represents the number of feature maps in the first block of DTNet.

To inspect the importance of the proposed structures, we conducted experiments progressively under different settings. The results are listed in Table III. DTNet performed well when compared with the linear transformation ATNet, showing the advantages of deformable transformation. Besides, when these two types of methods are used in combination, the performance is further improved, achieving a Dice score of 0.8258. These experiments prove that the proposed coarse-to-fine framework is effective. Utilizing this framework, we add the proposed JL constraint, and the Dice score is increased by another 0.8% thanks to the effective suppression of the wrongly predicted background pixels. Our final model is constructed by introducing the proposed dual consistency constraint, achieving the best Dice score of 0.8397.

We derived the Jacobian determinant to calculate the number of folding pixels (the lower the better) to check the model effectiveness. The coarse-to-fine framework, ATNet + DTNet, gets a number of 30 ± 30, which is significantly lower than that of DTNet (46 ± 50). However, this number is slightly increased when adding the JL constraint. We suspect that JL is designed for background error suppression, which might lead to the folding of unexpected background pixels. Nevertheless, with the introduction of the dual consistency constraint, the number is largely reduced to 13 ± 15. This proves that the dual consistency constraint can effectively suppress the occurrence of pixels folding through the inverse transformation.

As stated in the previous sections, the four weights in the JL function (Eq. 8), $\lambda_{i \in \{1,2,3,4\}}$, were empirically set to 1, 4, 100, 100. Here, we conducted experiments by fixing λ_1 and λ_4 to investigate the influence of λ_2 and λ_3, which are also the α and β in Eq. 6 that control the relative contributions of MI loss and the prior knowledge-based background suppressing loss. In details, we checked different α values from 0 to 10 with a step size of 1, and different β values from 0 to 200 with a step size of 20. The results are shown in Fig. 6. Two conclusions can be made. Firstly, with the increase of α, the registration performance gradually improves until the Dice scores fluctuate around 0.83. This indicates that MI

TABLE III
QUANTITATIVE RESULTS COMPARISON BETWEEN DIFFERENT METHODS.

Method	Dice	Precision	Recall	$\lvert J_\varphi\rvert < 0$
Undef	0.7822 ± 0.0974	0.8491 ± 0.1007	0.7326 ± 0.1175	–
ATNet	0.8067 ± 0.0935	0.8699 ± 0.0948	0.7590 ± 0.1128	–
DTNet	0.8117 ± 0.0919	0.8727 ± 0.0979	0.7690 ± 0.1173	46 ± 50
ATNet+DTNet	0.8258 ± 0.0787	0.8847 ± 0.0826	0.7905 ± 0.1112	30 ± 30
ATNet+DTNet(JL)	0.8339 ± 0.0840	$\mathbf{0.8889 \pm 0.0801}$	0.7955 ± 0.1165	60 ± 50
Ours	$\mathbf{0.8397 \pm 0.0756}$	0.8856 ± 0.0808	$\mathbf{0.8081 \pm 0.1069}$	13 ± 15

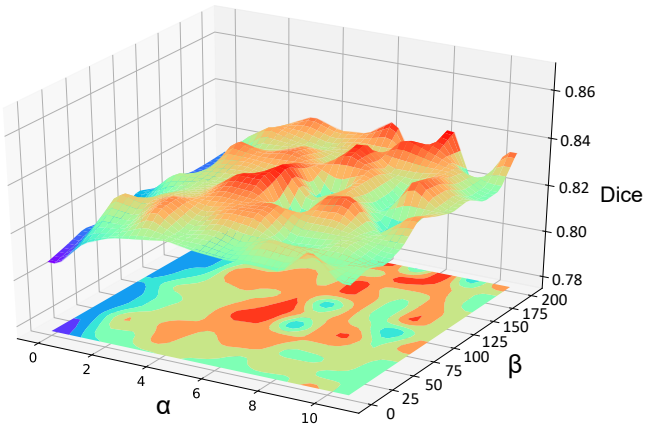

Fig. 6. Influence of the weights (α and β) in the proposed JL on the registration performance.

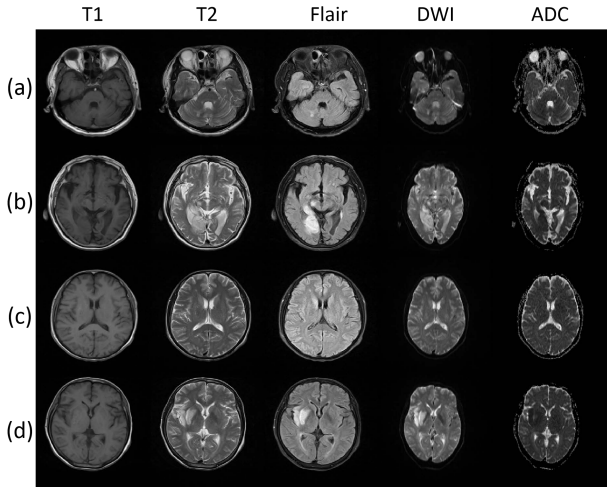

Fig. 7. MR multi-contrast brain images without physical alignment in advance. The images show large discrepancies between images acquired with different contrasts.

is important for accurate image registration. Secondly, with the increase of β, the registration performance also improves slightly. This confirms that the proposed prior knowledge-based background suppressing loss can help MI loss better optimize the network. The best Dice score of 0.8397 is achieved when $\alpha = 4$ and $\beta = 100$, which is much better than the Dice score of 0.8289 when $\alpha = 4$ and $\beta = 0$. Overall, the registration performance is quite robust with changing α and β values, and the proposed JL is effective. These results in all confirm that the proposed coarse-to-fine architecture, JL, and the dual consistency constraint can successfully enhance the multi-contrast MR image registration performance.

G. Experiment on data without scanning information

There are occasions when the scanning information, including the pixel spacing and field of views, is lost. Without the scanning information, multi-contrast images cannot be pre-aligned in the physical space, which brings great difficulties to the accurate image registration task. To increase the application capability of the proposed method, we conduct experiments without the first step physical space alignment. In this set of experiments, all the collected 555 image data are utilized. Some examples are shown in Fig. 7. It can be observed that due to the different imaging parameters, large discrepancies exist between the different contrast images.

The registration results of different methods are reported in Table IV. Overall, worse results are obtained in these experiments compared to those achieved with the physical space alignment (Table I). Before any registration, stroke lesion annotations in FLAIR and DWI images are largely misaligned with an average Dice score of only 0.3472. Compared to the scores achieved with the first step physical space alignment (Table I), the performance of SyN (Only) is dramatically decreased by more than 20% (0.5880 vs. 0.8101). SyN (Affine) obtains a slightly decreased score of 0.8048. The scores of all the learning-based comparison methods are decreased by roughly 4%. Our method still maintains a good performance with a score of 0.8260, which is only decreased by 1.37%. It indicates that the proposed method generalizes well to difficult tasks, and thus, the robustness is improved. Considering the time complexities, our proposed method becomes better than the time-consuming iterative-based method of SyN (Affine). Overall, when facing more challenging tasks, our method can still maintain good registration performance with satisfactory registration speed.

Moreover, we also tested to the structural MR images acquired with the three contrasts (T1 weighted, T2 weighted, and FLAIR) to DWI images using the proposed method (Fig. 8). Results indicate that our method can also perform quite well, which shows the general applicability of our method when handling different multi-contrast MR image registration tasks. It again validates that robustness of our method, and its high potential to be applied in clinical practices.

V. CONCLUSION

Multi-contrast MR image registration is critical for many clinical applications. Existing registration methods are limited

TABLE IV
QUANTITATIVE MEASUREMENT OF THE STROKE REGION REGISTRATION RESULTS AND THE REQUIRED TEST TIME BASED ON DATA WITHOUT PRE-ALIGNMENT.

Method	Dice	Precision	Recall	Sec/Slice (GPU)	Sec/Slice (CPU)
Undef	0.3472 ± 0.2390	0.3092 ± 0.2131	0.3999 ± 0.2789	-	-
SyN(Only)	0.5880 ± 0.2502	0.5539 ± 0.2628	0.6534 ± 0.2584	-	3.0238
SyN(Affine)	0.8084 ± 0.0989	$\mathbf{0.8682 \pm 0.1411}$	0.7565 ± 0.1510	-	3.7044
LT-Net	0.7599 ± 0.1371	0.8283 ± 0.1338	0.7178 ± 0.1597	0.0152	0.2242
VM-diff	0.7583 ± 0.1181	0.8166 ± 0.1326	0.7181 ± 0.1283	0.0401	0.1776
VM	0.7672 ± 0.1281	0.7984 ± 0.1358	0.7480 ± 0.1403	0.0109	0.1792
ATNet	0.7603 ± 0.1279	0.8066 ± 0.1307	0.7267 ± 0.1438	**0.0101**	**0.0351**
Ours	$\mathbf{0.8260 \pm 0.0761}$	0.8666 ± 0.0921	$\mathbf{0.7981 \pm 0.0989}$	0.0201	0.2176

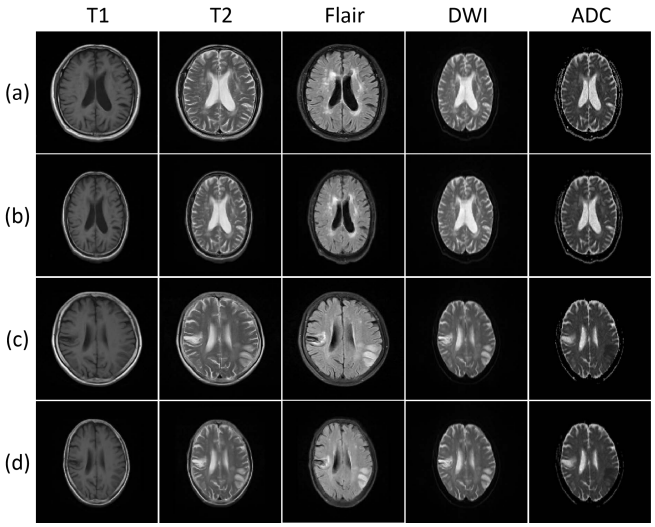

Fig. 8. Example results from registering MR images without pre-alignment acquired by the three structural sequences (T1 weighted, T2 weighted, and FLAIR) to DWI. (a) and (c) are two image slices selected from one patient without pre-alignment. (b) and (d) are the corresponding registration results.

by either the registration performance or the registration speed. In this paper, we propose a novel unsupervised deep learning-based registration framework. The proposed method embeds an affine transformation network in a deformable transformation network, which can not only improves the multi-contrast MR image registration performance but also reduces the time requirement for the registration process. In addition, a dual consistency strategy is proposed to achieve bi-directional image registrations so that the robustness of the method can be enhanced. To optimize the framework, we also developed a joint loss function combining the mutual information loss with an elaborately designed prior knowledge-based background suppressing loss. Compared to state-of-the-art registration methods, our framework achieves the best performance with a Dice score of 0.8397. Our method is also 10 times faster than the most competitive method (SyN) when testing on a CPU. In addition, our method can maintain the performance when handling different tasks, while comparison methods show large performance degradations.

Our developed method is not limited to multi-contrast MR image registrations. It can also be applied to unimodal or other multi-modal image registration tasks with modifications. Furthermore, accurate and efficient registration algorithms can be employed in the development of learning-based methods when human annotations are expensive to obtain and reduced reliance on annotations is necessary. For example, the proposed method can be easily extended to ATLAS-based segmentation tasks. In the future, we expect to further develop the proposed method to accommodate multi-modal image registrations such as those from CT to MR images. Overall, our method presents encouraging potentials in assisting intelligent medical data analysis.

VI. ACKNOWLEDGMENTS

This research was partly supported by Scientific and Technical Innovation 2030-"New Generation Artificial Intelligence" Project (2020AAA0104100, 2020AAA0104105), the National Natural Science Foundation of China (61871371, 81830056), Key-Area Research and Development Program of GuangDong Province (2018B010109009), the Basic Research Program of Shenzhen (JCYJ20180507182400762), Youth Innovation Promotion Association Program of Chinese Academy of Sciences (2019351).

REFERENCES

[1] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi, "Learning a probabilistic model for diffeomorphic registration," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2165–2176, 2019.

[2] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9252–9260.

[3] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 729–738.

[4] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[5] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical image analysis*, vol. 57, pp. 226–236, 2019.

[6] S. Wang, S. Tan, Y. Gao, Q. Liu, and D. Liang, "Learning joint-sparse codes for calibration-free parallel mr imaging (lindberg)," *IEEE Transactions on Medical Imaging*, vol. PP, no. 99, pp. 1–1, 2017.

[7] K. Chaisaowong and M. Jiang, "An automated 3d-atlas-based registration towards the anatomical segmentation of pulmonary pleural surface," in *2018 International ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI-NCON)*. IEEE, 2018, pp. 85–88.

[8] K. Marstal, F. Berendsen, N. Dekker, M. Staring, and S. Klein, "The continuous registration challenge: Evaluation-as-a-service for medical image registration algorithms," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1399–1402.

[9] X. Cao, J. Yang, Y. Gao, Q. Wang, and D. Shen, "Region-adaptive deformable registration of CT/MRI pelvic images via learning-based image synthesis," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3500–3512, 2018.

[10] X. Cao, J. Yang, J. Zhang, Q. Wang, P.-T. Yap, and D. Shen, "Deformable image registration using a cue-aware deep regression network," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1900–1911, 2018.

[11] Y. Zhao, S. Zhang, H. Chen, W. Zhang, J. Lv, X. Jiang, D. Shen, and T. Liu, "A novel framework for groupwise registration of fMRI images based on common functional networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 485–489.

[12] L. König, A. Derksen, M. Hallmann, and N. Papenberg, "Parallel and memory efficient multimodal image registration for radiotherapy using normalized gradient fields," in *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*. IEEE, 2015, pp. 734–738.

[13] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *information Fusion*, vol. 33, pp. 100–112, 2017.

[14] E. O. Stejskal and J. E. Tanner, "Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient," *The journal of chemical physics*, vol. 42, no. 1, pp. 288–292, 1965.

[15] P. J. Basser, J. Mattiello, and D. LeBihan, "Mr diffusion tensor spectroscopy and imaging," *Biophysical journal*, vol. 66, no. 1, pp. 259–267, 1994.

[16] P. C. Lauterbur, "Image formation by induced local interactions: examples employing nuclear magnetic resonance," *nature*, vol. 242, no. 5394, pp. 190–191, 1973.

[17] D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Physics in medicine & biology*, vol. 46, no. 3, p. R1, 2001.

[18] L. G. Brown, "A survey of image registration techniques," *ACM computing surveys (CSUR)*, vol. 24, no. 4, pp. 325–376, 1992.

[19] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens, "Automated model-based bias field correction of mr images of the brain," *IEEE transactions on medical imaging*, vol. 18, no. 10, pp. 885–896, 1999.

[20] B. M. Dawant, "Non-rigid registration of medical images: purpose and methods, a short survey," in *Proceedings IEEE International Symposium on Biomedical Imaging*. IEEE, 2002, pp. 465–468.

[21] T. G. Reese, O. Heid, R. Weisskoff, and V. Wedeen, "Reduction of eddy-current-induced distortion in diffusion mri using a twice-refocused spin echo," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 49, no. 1, pp. 177–182, 2003.

[22] R. Bajcsy and S. Kovačič, "Multiresolution elastic matching," *Computer vision, graphics, and image processing*, vol. 46, no. 1, pp. 1–21, 1989.

[23] D. Shen and C. Davatzikos, "Hammer: hierarchical attribute matching mechanism for elastic registration," *IEEE transactions on medical imaging*, vol. 21, no. 11, pp. 1421–1439, 2002.

[24] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International journal of computer vision*, vol. 61, no. 2, pp. 139–157, 2005.

[25] G. L. Hart, C. Zach, and M. Niethammer, "An optimal control approach for deformable registration," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2009, pp. 9–16.

[26] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[27] Z. Chen, H. Jin, Z. Lin, S. Cohen, and Y. Wu, "Large displacement optical flow from nearest neighbor fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2443–2450.

[28] J. Wulff and M. J. Black, "Efficient sparse-to-dense optical flow estimation using a learned basis and layers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 120–130.

[29] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast mr images," *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[30] J.-P. Thirion, "Image matching as a diffusion process: an analogy with Maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, p. 243–260, 1998.

[31] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, "Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy," *Physics in Medicine & Biology*, vol. 50, no. 12, p. 2887, 2005.

[32] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Non-parametric diffeomorphic image registration with the demons algorithm," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2007, pp. 319–326.

[33] Z. Shen, F.-X. Vialard, and M. Niethammer, "Region-specific diffeomorphic metric mapping," in *Advances in Neural Information Processing Systems*, 2019, pp. 1098–1108.

[34] Z. Shen, X. Han, Z. Xu, and M. Niethammer, "Networks for joint affine and non-parametric image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4224–4233.

[35] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[36] J. Krebs, T. Mansi, B. Mailhé, N. Ayache, and H. Delingette, "Learning structured deformations using diffeomorphic registration," *arXiv preprint arXiv:1804.07172*, 2018.

[37] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. K. Maier, N. Ayache, R. Liao, and A. Kamen, "Robust non-rigid registration through agent-based action learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 344–352.

[38] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, and D. Comaniciu, "An artificial agent for robust image registration," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[39] K. Ma, J. Wang, V. Singh, B. Tamersoy, Y.-J. Chang, A. Wimmer, and T. Chen, "Multimodal image registration with deep context reinforcement learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 240–248.

[40] S. Miao, S. Piat, P. Fischer, A. Tuysuzoglu, P. Mewes, T. Mansi, and R. Liao, "Dilated fcn for multi-agent 2d/3d medical image registration," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[41] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical image analysis*, vol. 52, pp. 128–143, 2019.

[42] H. Li and Y. Fan, "Non-rigid image registration using self-supervised fully convolutional networks without training data," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1075–1078.

[43] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, p. 8, 2020.

[44] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2017–2025. [Online]. Available: http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf

[45] I. Yoo, D. G. Hildebrand, W. F. Tobin, W.-C. A. Lee, and W.-K. Jeong, "ssemnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 249–257.

[46] J. Krebs, T. Mansi, B. Mailhé, N. Ayache, and H. Delingette, "Unsupervised probabilistic deformation modeling for robust diffeomorphic registration," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 101–109.

[47] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[48] Z. Li, F. Huang, J. Zhang, B. Dashtbozorg, S. Abbasi-Sureshjani, Y. Sun, X. Long, Q. Yu, B. ter Haar Romeny, and T. Tan, "Multi-modal and multi-vendor retina image registration," *Biomedical optics express*, vol. 9, no. 2, pp. 410–422, 2018.

[49] J. Ceranka, M. Polfliet, F. Lecouvet, N. Michoux, J. de Mey, and J. Vandemeulebroucke, "Registration strategies for multi-modal whole-body MRI mosaicing," *Magnetic resonance in medicine*, vol. 79, no. 3, pp. 1684–1695, 2018.

[50] X. Cao, J. Yang, Y. Gao, Y. Guo, G. Wu, and D. Shen, "Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis," *Medical image analysis*, vol. 41, pp. 18–31, 2017.

[51] X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, and D. Shen, "Deep learning based inter-modality image registration supervised by intra-modality similarity," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2018, pp. 55–63.

[52] P. Dong and N. P. Galatsanos, "Affine transformation resistant watermarking based on image normalization," in *Proceedings. International Conference on Image Processing*, vol. 3. IEEE, 2002, pp. 489–492.

[53] S. Wang, S. Cao, D. Wei, R. Wang, K. Ma, L. Wang, D. Meng, and Y. Zheng, "Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9162–9171.

[54] I. Vajda, *Theory of statistical inference and information*. Kluwer Academic Pub, 1989, vol. 11.

[55] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[56] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.