A coarse-to-fine framework for unsupervised multi-contrast MR image deformable 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 essential 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 get rid of both the multi-step iteration process and the complex image preprocessing operations. 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 that consists of 555 cases, with encouraging performances achieved. Compared to the commonly utilized registration methods, including Voxelmorph, SyN, and LDDMM, the proposed method achieves the best registration performance with a Dice score of 0.826 in identifying stroke lesions. More robust performance in low-signal areas is also observed. With regards to the registration speed, our method is about 17 times faster than the most competitive method of SyN when testing on a same CPU.

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

I. INTRODUCTION

Multi-modal medical imaging plays an important role in many clinical applications [1-11], such as image-guided intervention, disease diagnosis, and treatment planning. Among them, multi-contrast magnetic resonance (MR) imaging is one of the most prevalent techniques utilized in brain imaging as different MR imaging sequences can highlight different regions of interest. However, multi-contrast MR image interpretation by the radiologists is time-consuming for careful comparison and sometimes difficult due to the following three reasons. Firstly, different image properties exist due to the different parameter settings of MR imaging sequences, including image resolution, slice thickness, and field of view (FOV). Secondly, inevitable physiological activity may lead to misalignment between the acquired multi-contrast MR images. Lastly, different imaging sequences generate MR images with different intensity distributions, which elevates the difficulties of manual reading of the images. Three examples are shown in Figure 1. It is obvious that images obtained with different sequences have varied morphologies, and the brain lesions annotated according to the FLAIR images are different from those annotated according to the DWI images. As a result, multi-contrast MR image registration is needed.

Fig.1. MR multi-contrast brain images acquired from three candidates. ‘Flair Label’ and ‘DWI Label’ represent strokes annotated by clinicians based on FLAIR and DWI respectively.

Different image registration methods are available. 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 which can speed up the registration process but at the cost 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. The affine transformation network is embedded in the deformable transformation network and end-to-end image registration is realized.

2) A dual consistency constraint is designed to maximize the cross-correlation in the space of topology maps of multi-contrast images. The reverse deformation field is derived and bi-directional deformations are achieved. The designed consistency constraint is enforced on the bi-directional deformations, and the robustness of the model generated transformation field can be better guaranteed.

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) We instantiate the proposed registration framework with multi-contrast MR images. Extensive experiments verify the effectiveness of the proposed model.

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 [12], [13], fluid [14]-[18] or B-spline models [19], are usually based on the iterative numerical solution of the optimization problem. Especially, in 1998, Thirion et al. [20] 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. [21], [22]. In 2005, Beg et al. [14] proposed another famous registration algorithm, LDDMM (Large Displacement Diffeo-morphic Metric Mapping), by deducting and implementing the Euler-Lagrange optimization to compute particle flows, solving a global variational problem, and estimating metrics for images. Subsequently, variants of this algorithm were also proposed, including RDMM, vSVF, and SYN [23]-[25]. Among them, SYN [25] 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], [26].

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 [27]-[30]. With the increased demand on the registration speed, single deep learning-based registration methods were proposed [2], [31]-[33]. 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 [34]. Yoo et al. [35] 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. [26], [36] proposed a random latent space learning method to alleviate the requirement on spatial regulariztion. De Vos et al. [31] 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 [37]. Li et al. [38] 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. [39] 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. [40]
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 $F \supset \{f^1, f^2, \ldots, f^K\}$ and $M \supset \{m^1, m^2, \ldots, m^K\}$ refer to the paired fixed image sets and moving image sets.

A. Affine transformation network – ATNet

STN [34] 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 [42], which geometrically consists of a non-singular linear transformation (transformation using a linear function). To clearly demonstrate the procedure, let $p(x_i, y_i)$ 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:

$$A_p(x) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$$  \hspace{1cm} (1)

where $\theta$ 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 $M_A \supset \{m^1_A, m^2_A, \ldots, m^K_A\}$. These predictions are roughly aligned to $M$, and dense deformation transformations are needed to align the detailed local structures. It is obvious that only performing a linear transformation will not be able to capture the subtle differences between multi-contrast images. Besides, affine transformations are global image transformations, which may lead to compromised predictions in regions with low signals. 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 $\varphi$ as the obtained transformable field. Each value in $\varphi$ represents an offset distance. Symbol $\circ$ refers to the transformation operator for $m$, which consists of pixel shifting and interpolation. For each pixel $p$ in $m$, pixel transformation can be defined as:

$$p' = p + \varphi(p)$$  \hspace{1cm} (2)

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

$$m \circ \varphi(p) = \sum_{q \in Z(p)} m(q) \prod_{d \in \{x, y\}} (1 - |p_{dim} - q_{dim}|)$$  \hspace{1cm} (3)

where $Z$ represents the regions composed of eight neighboring 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. A loss function that can measure pixel-wise similarity is generally needed to supervise the network optimization (such as NCC and mean square error (MSE)), especially when large scale deformable transformations are performed. Nonetheless, it is difficult to construct an appropriate constraint for multi-contrast image registration where big differences between images exist.

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

To reduce the challenges of large scale image transformation, we propose a coarse-to-fine image registration framework. Specifically, we embed the pre-trained ATNet, $D_p(F,M)$, 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 conventional registration method, our method does not need to iterate over affine or deformation transformations [25]. Moreover, compared with the existing deep learning-based registration method, our method can realize online image affine transformations without the need for image preprocessing [2]-[5]. Meanwhile, we can directly 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 [25] with an Euler Lagrange equation for iterative optimization and achieved great success in medical image registration. Inspired by this work, we propose our bi-directional image transformation method.

As defined in the previous section, \( \varphi \) is the transformable field for the forward transformation of registering moving images to fixed images. Accordingly, we constructed the reverse deformation field \( \varphi^{-1} \) to guide the backward transformation of registering fixed images to moving images. Instead of building a new network to generate \( \varphi^{-1} \) from scratch [43], we directly generate \( \varphi^{-1} \) from \( \varphi \). In details, we first decompose \( \varphi \) to obtain the horizontal and vertical offset fields respectively. Then, we warp both offset fields with \( \varphi \) to form deformed offset fields, which can match the pixel of the original predictions. By recombining the deformed offset fields, a new transformable field is generated. Finally, the reverse transformation field \( \varphi^{-1} \) is obtained by multiplying with -1.

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 of the whole framework \( (M_D \supset \{m_D^1, m_D^2 \ldots m_D^d \}) \) instead of the fixed images \( F \) to calculate the reverse transformed images \( M_D^{-1} \):

\[
M_D^{-1} = M_D \circ \varphi^{-1} = M_D \circ (-\sum_{d=1}^{D} (\varphi_{d,im} \circ \varphi))
\]

We assume that \( M_D^{-1} \) should still maintain the same distribution as \( M_A \). Based 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

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\( \theta \)) 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 transformable field \( \varphi \). With \( \varphi \), the final prediction M\( C \) is generated. 3) A dual consistency constraint. We propose a novel reverse transformation from M\( D \) to M\( D^{-1} \) to further enhance the registration performance. We calculated the reverse transformable field \( \varphi^{-1} \) and warp it with M\( D \) 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 \( L_k(\cdot) \) to represent an undefined network which can be either ATNet (A\( \theta (\cdot) \)) or DTNet (D\( \theta (\cdot) \)).

The most important loss function used is Mutual Information (MI), which can measure the distribution dependence between two random variables [37]. Here, we define two marginal probability distributions, \( p_f(f) \) and \( p_M(m) \), and a joint probability distribution obtained from F and M, \( p_{FM}(f, m) \). MI (MI(F,M)) measures the degree of dependence between F and M by measuring the distance between the joint distribution \( p_{FM}(f, m) \) and the distribution associated with the case of complete independence \( (p_f(f) \cdot p_M(m)) \) by means of the Kullback-Leibler measurement [44], as shown in Eq.5:

\[
\text{MI}(F, M) = \sum_{f, m} p_{FM}(f, m) \log \frac{p_{FM}(f, m) \cdot p_{f}(f) \cdot p_{M}(m)}{p_{f}(f) \cdot p_{M}(m)}
\]

If \( F \) and \( \xi_{\theta}(F, M) \) are independent, \( p_{FM}(f, m) = p_f(f) \cdot p_M(m) \), and MI(F, \xi_{\theta}(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 [37].

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 \( L(F, \mathcal{X}_0(F), M, \lambda) \):

\[
\begin{align*}
L(F, \mathcal{X}_0(F), M, \lambda) &= \sum_{f,m} (\alpha M1(f, \mathcal{X}_0(f, m))) + \beta \sum_i (ME(f_i, \mathcal{X}_0(f, m_i)), if \ f_i < \gamma \ 0, otherwise)
\end{align*}
\]

where \( i \in \mathbb{N} \) represents the pixels in images, \( \gamma \) is a threshold obtained from the data set to determine whether the pixel is background or not, \( \alpha \) and \( \beta \) 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 \( M_0^{-1} \) and \( M_0 \). The utilization of MSE loss is not fixed and can be replaced by similar losses, such as NCC or L1 norm.

The last loss function is to constrain the transformable field \( \varphi \). 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 transformable field \( \varphi \):

\[
SL(\varphi) = \sum_{f,m} (\nabla \varphi(f, m))^2
\]

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

The overall loss function to optimize the framework is calculated as follow:

\[
\text{Loss}_{\varphi}(F, M) = \lambda_1 SL(\varphi) + L(F, D_\theta(F, M), \lambda_2, \lambda_3) + \lambda_4 M3E(\lambda_0(F, M), D_\theta^{-1}(F, M))
\]

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 DWT are the most commonly used MR weighted sequences. Thus, our image registration experiments are mainly conducted with Flair and DWT data.

A. Dataset

The multi-contrast MR data were collected by Guizhou Provincial People's Hospital. 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 DWT. All images were obtained with a Siemens 1.5T scanner. It is worth noting that the scan parameters of the five sequences are not fixed and the obtained images are not perfectly aligned (Fig. 1). As an example, the FOV phase amplitude for DWI is 83.5%-100%, while for FLAIR, it is only 70%-87.5%. The slice thickness ranges from 6.0 mm to 7.0 mm and the slice interval ranges from 7.8 mm to 9.1 mm. Among the 555 cases, 40 cases were randomly selected and the stroke lesions in DWI and FLAIR images were annotated by experienced clinicians using ITK-snap. These 40 annotated cases are treated as the test set with the rest 515 cases as 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 very good image 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 588k trainable parameters.

DTNet is modified from the famous UNet with an encoder-decoder architecture [45]. 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 transformable field \( \varphi \) can be obtained. DTNet has about 1478k trainable parameters.

We adopt Symmetric Normalization (SyN) [25], the famous top-performing brain registration algorithm, as one baseline for comparison. It is implemented in the publicly available Advanced Normalization Tools (ANTS) software package [46] with a MI constraint for multi-contrast MR image registration. In our implementation, SyN has three designs: 1) Moving images go through ANTs-based affine transformations and SyN, represented as ‘SyN(Affine)’; 2) Moving images go through ATNet and then SyN, represented as ‘SyN(Affine Net)’; 3) Moving images go through SyN only, represented as ‘SyN(Only)’. For the second baseline, we choose the LDDMM algorithm, which is implemented at https://github.com/SteffenCzolbe/pyLDDMM with 50 iterations. 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. 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 50 minutes, and the entire framework including DTNet requires another 20 minutes to optimize. The four weights in the loss function, $\lambda_{\text{it}}\{1, 2, 3, 4\}$, were set to $\{1, 7, 160, 100\}$ empirically. The threshold factors $\gamma$ in the JL was set to 0.1. Our code will be available online at https://github.com/SZUHvern.

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 find that the use of the affine transformation can obviously enhance the registration performance SyN (compare the results of SyN with and without affine transformation). Between the ATNet-based SyN (SyN (Affine)) and the ANTs toolkit-based SyN (SyN (Affine)), no obvious difference is observed in the registered results, which indicate that the ATNet achieves effective affine transformations. VoxelMorph predictions are morphologically similar to our proposed method, but there are more artifacts and blurred regions, which may be caused by unconstrained large-scale pixel migration. The conventional iterative algorithm LDDMM show large artifacts in low-intensity areas (such as the skulls) and the overall performance is unsatisfactory due to image distortion. In summary, qualitative results confirm that our proposed method performs better than or at least similar to the state-of-the-art methods for multi-contrast MR image registration.

The registration performance of stroke lesions is quantified by calculating three metrics, the Dice score, precision, and recall, between the registration results and the corresponding fixed DWI images. The results together with the testing time are listed in Table 1. Boxplots of Dice scores for each method are shown in Fig.4.

| Method         | DICE (Mean ± SD) | Recall (Mean ± SD) | Sec/Sl (CPU) |
|----------------|------------------|--------------------|--------------|
| Undef          | 0.347 ± 0.239    | 0.400 ± 0.279      | 425.200      |
| LDDMM          | 0.580 ± 0.265    | 0.560 ± 0.221      | 3.024        |
| SyN(Affine)    | 0.808 ± 0.099    | 0.757 ± 0.153      | 3.704        |
| SyN(Affine) (Undef) | 0.808 ± 0.120 | 0.757 ± 0.120      | 2.396 ± 0.035 |
| Afline Net     | 0.790 ± 0.128    | 0.727 ± 0.144      | 0.010        |
| VoxelMorph     | 0.707 ± 0.138    | 0.746 ± 0.136      | 0.011        |
| Ours           | 0.820 ± 0.075    | 0.785 ± 0.099      | 0.020        |

Example predictions of different methods are shown in Fig.3. The first two columns present the moving FLAIR images and the fixed DWI images with the annotated stroke lesions. The third columns show the registration results of our proposed method, and the rest columns are the results generated by the comparison algorithms. Overall, all the methods show satisfactory registration results when considering the stroke lesions with only several exceptions. From example (a), it is clear that our method gives better registration results showing by the higher overlap between the green and blue lines. Comparable results are obtained by our method and SYN (Affine) for examples (b-d). We suspect that the good registration results in the stroke lesion regions may be caused by the high-intensity signals, which makes it relatively easier to be accurately registered. Nevertheless, for other anatomical structures, especially in low-intensity areas, our method can still maintain good registration performance when other methods fail to do so. For example, undesirable deformations are observed from the registration results with the red arrow of the structures of eyes or cerebrospinal fluid (Fig.3 b-d).

![Fig.3. Qualitative registration results of different methods. The blue and red lines indicate the stroke lesion regions annotated by radiologists based on DWI and Flair, and the green lines indicate the predictions of different methods. Because some of the lesions are too small, we enlarge the regions and place them in the middle.](image)

![Fig.4. Boxplots of registration results comparing the undeformed (Undef) case to the different algorithms: LDDMM, Aflne Net, VoxelMorph, SyN(only), SyN(Affine Net), SyN(Affine) and Ours. Structures are ordered by mean Dice score.](image)
the necessary of multi-contrast image registration. SyN (Only) and LDDMM achieve unsatisfactory registration results that only slightly increase the Dice scores. Introducing the affine transformation, the performance of SyN was significantly improved. For the two different implements of affine transformation before SyN, the Dice score of SyN (Affine Net) is relatively lower than that of SyN (Affine) but within an acceptable range. Compared to all the comparison methods, our method achieves the best Dice score of 0.826, which again confirms the effectiveness of the proposed method.

Besides, the time efficiency of each method is also listed in Table 1. To conduct fair comparisons between the different methods, we use the required CPU time. Among all the methods, LDDMM is the slowest method that needs 425 seconds to register one image slice, and ATNet alone is the fastest method which has a registration speed of 0.01 seconds per slice. Comparing with the most competitive SYN (Affine) method, our method is about 17 times faster with better registration results. For one 3D medical image, say 20 slices, the proposed method can process it within 5 seconds, whereas SyN needs nearly 75 seconds. This indicates that our method is more applicable to clinical applications where real-time diagnosis is important.

![Image](image_url)

**Table 1**: Performance comparison of different methods.

| Method      | Dice Score | Time (sec) |
|-------------|------------|------------|
| SyN (Only)  | 0.75       | 425        |
| LDDMM       | 0.79       | 425        |
| Syn (Affine) | 0.826      | 5          |

**Fig. 5**: Example results from registering MR images acquired by the four sequences (T1 weighted, T2 weighted, FLAIR, and ADC) to DWI. (a) and (c) are two image slices selected from one patient before registration. (b) and (d) are the corresponding registration results.

In addition, we register the MR images acquired with all the four sequences to DWI images using the proposed algorithm to investigate the robustness of the method. Results are shown in Fig. 5. In our collected dataset, there are five types of images, namely, DWI, ADC, T1 weighted, T2 weighted and FLAIR sequences. Our method can handle all the registration tasks without obvious performance compromise. This verifies the robustness of our proposed method, which can be readily applied to different tasks where multi-contrast image registration is in need.

**D. Analysis on the transformable field**

Visualizations of the transformable fields $\varphi$ are shown in Fig. 6. It gives a direct measurement of the pixel displacement during image registration. Regions with large deformations are highlighted in red circles in our registration results. More specifically, we find these large deformations result in the following scenarios: the right side of cerebellum bulges upward (slice a), the bilateral cerebrospinal fluid is more concentrated (slice b), and the frontal lobe is offset downward (slice c). These large local deformations all contribute to more accurate registration results.

**Fig. 6**: Visualizations of the transformable fields $\varphi$. 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.

**E. 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 proposed joint loss function JL. Finally, we check the necessities of the embedded ATNet and the dual consistency constraint.

In Fig. 7, 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. A smaller learning rate might be more appropriate. For the different network widths, significantly worse performance is observed with a width of 8, which might indicate that the network is not able to capture the complex image properties. Wider networks with widths of 16, 32, and 64 show similar performance. 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.
IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. x, NO. x, 2020

To inspect the importance of the proposed joint loss function, we conducted experiments with two loss functions, JL and MI, under different network settings (Table 2). In all correspondingly paired experiments (VoxelMorph JL vs. VoxelMorph MI, Affine Net JL vs. Affine Net MI, and Ours JL vs. Ours MI), JL consistently improves the registration performance. This phenomenon indicates that in multi-contrast MR image registration tasks, utilizing the MI constraint solely is not enough. Prior regarding the medical images should be introduced to achieve better performance.

| Method       | DICE   | Precision | Recall  |
|--------------|--------|-----------|---------|
| VM (MI)      | 0.767 ± 0.128 | 0.798 ± 0.136 | 0.748 ± 0.140 |
| VM (JL)      | 0.778 ± 0.127 | 0.827 ± 0.130 | 0.744 ± 0.142 |
| VM (JL)+Affine | 0.810 ± 0.090 | 0.867 ± 0.095 | 0.774 ± 0.119 |
| VM (JL)+Dual | 0.718 ± 0.147 | 0.745 ± 0.147 | 0.703 ± 0.161 |
| Affine Net (MI) | 0.760 ± 0.128 | 0.807 ± 0.131 | 0.727 ± 0.144 |
| Affine Net (JL) | 0.708 ± 0.107 | 0.825 ± 0.107 | 0.729 ± 0.124 |
| Ours (MI)    | 0.815 ± 0.009 | 0.854 ± 0.100 | 0.773 ± 0.109 |
| Ours (JL)    | 0.826 ± 0.076 | 0.867 ± 0.092 | 0.798 ± 0.099 |

As stated in the previous sections, the four weights in the JL function (eq. 8), $\lambda_i \in \{1, 7, 16, 100\}$, were empirically set to $\{1, 7, 16, 100\}$. Here, we conducted experiments by fixing $\lambda_1$ and $\lambda_2$ to investigate the influence of $\lambda_3$ and $\lambda_4$, which are also the $\alpha$ and $\beta$ in eq. 6 that controls the relative contribution of MI loss and the prior knowledge-based background suppressing loss. In details, we checked different $\alpha$ values from 0 to 10 with a step size of 1, and different $\beta$ values from 0 to 200 with a step size of 20. The results are shown in Fig.8. Two conclusions can be made. Firstly, with the increase of $\alpha$, the registration performance gradually improves until the Dice scores fluctuate around 0.81. This indicates that MI is important for accurate image registration. Secondly, with the increase of $\beta$, 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 0.826 is achieved when $\alpha = 7$ and $\beta = 160$, which is much better than the Dice score of 0.815 when $\alpha = 7$ and $\beta = 0$. Overall, the registration performance is quite robust with changing $\alpha$ and $\beta$ values, and the proposed JL is effective.

Results relevant to the necessities of the embedded ATNet and the dual consistency constraint are also shown in Table 2. In these experiments, we treated VoxelMorph (VoxelMorph (JL)) as the baseline. When ATNet is embedded (VoxelMorph (JL) + Affine), the performance of VoxelMorph is significantly improved showing by an increase in the Dice score from 0.778 to 0.811. This reflects that the affine transformation can reduce the registration difficulty of deformable transformation models and better registration results can be obtained. However, when the dual consistency constraint in enabled (VoxelMorph (JL) + Dual), severely worse registration results are obtained. We suspect that the large-scale deformation of VoxelMorph makes it too difficult to restore the deformed images. Nevertheless, when comparing the results achieved by VoxelMorph with affine transformations (VoxelMorph (JL) + Affine) with those achieved by our method with the consistency constraint (Ours (JL)), obvious performance improvement is observed. These results confirm that with our framework, both the affine transformations and the dual consistency constraint can successfully enhance the multi-contrast MR image registration performance.

![Fig.7. 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.](image1)

![Fig.8. Influence of the weights ($\alpha$ and $\beta$) in the proposed JL on the registration performance.](image2)

V. Conclusion

Multi-contrast MR image registration is critical for many clinical applications. Existing registration methods are limited by either the registration performance or the registration speed. In this paper, we propose a novel unsupervised deep learning-based concise registration framework. The proposed method embed an affine transformation network in a deformable transformation network, which can not only improve the multi-contrast MR image registration performance but also greatly reduce 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 registration performance with a Dice score of 0.826 and a registration speed 17 times faster than the most competitive method (SyN) when testing on a single CPU.
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.

ACKNOWLEDGMENTS

This research was partly supported by the National Natural Science Foundation of China (61871371, 81830056), Key-Area Research and Development Program of GuangDong Province (2018B010109009), Science and Technology Planning Project of Guangdong Province (2017B02027012), the Basic Research Program of Shenzhen (JCZY20180507182400762), Youth Innovation Promotion Association Program of Chinese Academy of Sciences (2019351), Major projects of the Ministry of Science and Technology (2019AAA0104100, 2019AAA0104105).

REFERENCES

[1] J. Krebs, H. Delingette, B. Mailhe, 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] 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 (ECTINCON). IEEE, 2018, pp. 85–88.

[7] 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.

[8] 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.

[9] 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.

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

[11] L. Konig, 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.

[12] R. Bajcsy and S. Kovacic, “Multiresolution elastic matching,” in Computer vision, graphics, and image processing, vol. 46, no. 1, pp. 1–21, 1989.

[13] 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.

[14] M. F. Beg, M. I. Miller, A. Trouve, and L. Younes, “Computing large ’ deformation metric maps using geodesic flows of diffeomorphisms,” International journal of computer vision, vol. 61, no. 2, pp. 139–157, 2005.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] J.-P. Thirion, “Image matching as a diffusion process: an analogy with Maxwell’s demons,” Medical Image Analysis, vol. 2, no. 3, pp. 243–260, 1998.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.
[26] J. Krebs, T. Mansi, B. Mailhe, N. Ayache, and H. Delingette, “Learning structured deformations using diffeomorphic registration,” arXiv preprint arXiv:1804.07172, 2018.

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

[28] 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.

[29] K. Ma, J. Wang, V. Singh, B. Tamosy, 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.

[30] S. Miao, S. Piat, P. Fischer, A. Tuyuzoglu, P. Mewes, T. Mansi, and R. Liao, “Dilated fnn for multi-agent 2d/3d medical image registration,” in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

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

[32] 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.

[33] 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.

[34] 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

[35] J. 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.

[36] J. Krebs, T. Mansi, B. Mailhe, 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.

[37] 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.

[38] Z. Li, F. Huang, J. Zhang, B. Dashthbozorg, 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.

[39] J. Ceranka, M. Pofflet, F. Lecouvet, N. Michoux, J. de Mey, and J. Vandemeulebroecke, “Registration strategies for multi-modal wholebody MRI mosaicing,” Magnetic resonance in medicine, vol. 79, no. 3, pp. 1684–1695, 2018.

[40] 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.

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

[42] 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.

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

[44] I. Vajda, Theory of statistical inference and information. Kluwer Academic Pub, 1989, vol. 11. [45] 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.

[46] 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.