

Inverse-Consistent Deep Networks for Unsupervised Deformable Image Registration

Jun Zhang

Abstract—Deformable image registration is a fundamental task in medical image analysis, aiming to establish a dense and non-linear correspondence between a pair of images. Previous deep-learning studies usually employ supervised neural networks to directly learn the spatial transformation from one image to another, requiring task-specific ground-truth registration for model training. Due to the difficulty in collecting precise ground-truth registration, implementation of these supervised methods is practically challenging. Although several unsupervised networks have been recently developed, these methods usually ignore the inherent inverse-consistent property (essential for diffeomorphic mapping) of transformations between a pair of images. Also, existing approaches usually encourage the to-be-estimated transformation to be locally smooth via a smoothness constraint only, which could not completely avoid folding (typically means registration errors) in the resulting transformation. To this end, we propose an *Inverse-Consistent deep Network (ICNet)* for unsupervised deformable image registration. Specifically, we develop an *inverse-consistent constraint* to encourage that a pair of images are symmetrically deformed toward one another, until both warped images are matched. Besides using the conventional smoothness constraint, we also propose an *anti-folding constraint* to further avoid folding in the transformation. The proposed method does not require any supervision information, while encouraging the diffeomorphic property of the transformation via the proposed inverse-consistent and anti-folding constraints. We evaluate our method on T1-weighted brain magnetic resonance imaging (MRI) scans for tissue segmentation and anatomical landmark detection, with results demonstrating the superior performance of our ICNet over several state-of-the-art approaches for deformable image registration. Our code will be made publicly available.

Index Terms—Deformable image registration, deep network, inverse-consistent, anti-folding, brain MRI.

1 INTRODUCTION

As a fundamental task in medical image analysis, deformable image registration aims to establish dense, non-linear spatial correspondences between a pair of images (denoted as the source/moving image and the target/fixed image) [1]. For example, while it is difficult to compare brain magnetic resonance imaging (MRI) scans of different subjects due to significant anatomical variability [2], [3], deformable registration enables direct comparison of anatomical structures across scans, and thus, is crucial for understanding variability across populations and the longitudinal evolution of brain anatomy for individuals with brain diseases [4]–[6].

Over the past few decades, a variety of non-learning-based deformable registration methods have been proposed for medical image analysis [6]–[9]. Typically, these methods iteratively optimize a similarity function for each pair of images to non-linearly align voxels with the similar appearance, while encouraging local smoothness on the registration mapping [1]. Since the similarity function needs to be optimized from scratch for each pair of unseen images, *i.e.*, the inherent registration patterns shared across different images are ignored, these methods are usually slow to perform deformable registration in practical applications. To address these issues, supervised learning-based and deep-learning-based approaches have been developed for deformable image registration [10]–[12]. These methods typically rely on *task-specific ground-truth registration* to train a regression model for image registration. However, the difficulty of

collecting ground-truth information often limits their utility in real-world applications. In addition, the mapping learned by these supervised methods might be biased by the selected ground-truth registration.

Recently, several unsupervised deep-learning methods have been proposed for medical image registration [13]–[17] without using any pre-defined supervision information for network training, thus maintaining the unsupervised nature of deformable registration. Although these methods achieve better registration performance in comparison to traditional supervised-learning-based methods, the output transformations (*e.g.*, displacement field or flow) are usually asymmetric, *i.e.*, the inherent inverse-consistent property of transformations between a pair of images is ignored. Here, the inverse-consistent property means that to-be-learned optimal transformations would encourage that a pair of images are symmetrically deformed toward each other, and the two bidirectionally deformed images are finally matched. Unfortunately, previous studies usually independently estimate the transformation from an image **A** to an image **B** or from **B** to **A**, thus failing to ensure these transformations be inverse mappings for each other. Besides, most of the existing (supervised or unsupervised) algorithms utilize solely a spatial smoothness penalty to constrain the transformation, which could not completely avoid foldings (typically indicating errors) in the registration mapping. As an illustration, in Fig. 1, we show two flows generated by a state-of-the-art deep-learning method [17] trained with different (*i.e.*, strong vs. weak) contributions from the smoothness constraint. As shown in Fig. 1 (a), if we excessively encourage local smoothness of the to-be-estimated flow by using a large weight for the smoothness constraint, the

J. Zhang is with the Tencent AI Medical Center, China. (Email: xdzhangjun@gmail.com).

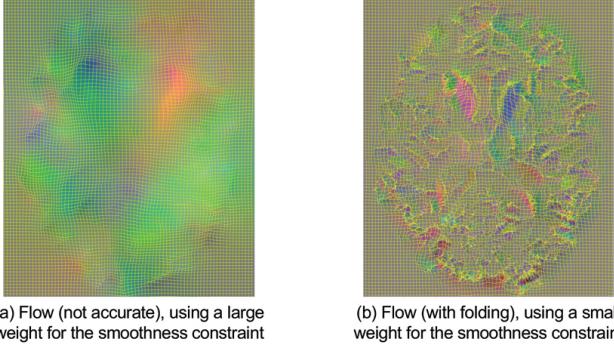


Fig. 1. Illustration of two flows learned by a state-of-the-art deep-learning method [17] using (a) a large weight for the smoothness constraint and (b) a small weight for the smoothness constraint, respectively.

obtained registration will be inaccurate due to global errors. Otherwise, there will be a lot of foldings in the learned flow (see Fig. 1 (b)), thus generating wrong registration due to local defects. That is, it is challenging to properly tune the contribution of the smoothness constraint to simultaneously avoid foldings in the estimated flow and maintain high registration accuracy.

To address these two issues, in this paper, we propose an Inverse-Consistent deep neural Network (ICNet) for unsupervised deformable image registration. Specifically, in ICNet, we develop an *inverse-consistent constraint* to encourage that a pair of images are symmetrically deformed toward each other in multiple passes, until the bidirectionally deformed images are matched to achieve correct registration. Besides using the conventional smoothness constraint, we also develop an *anti-folding constraint* to avoid foldings in the to-be-estimated flow. The proposed ICNet method does not require any supervision information, which also encourages the diffeomorphic property of the transformations via the proposed inverse-consistent and anti-folding constraints. We evaluate our method in both tasks of tissue segmentation and anatomical landmark detection with 3D T1-weighted brain MRI scans. The experimental results demonstrate the superior performance of the proposed method over several state-of-the-art methods in deformable image registration.

The rest of this paper is organized as follows. We briefly introduce relevant studies in Section 2. In Section 3, we present the proposed network, inverse-consistent constraint, and anti-folding constraint in detail. We then describe studied materials, competing methods, experimental settings and results in Section 4. We further analyze the influence of several essential strategies used in the proposed method in Section 5. We finally conclude this paper in Section 6.

2 RELATED WORK

2.1 Deformable Image Registration

Deformable image registration refers to an non-linear process of revealing the voxel-wise spatial correspondence between source and target images. Let \mathbf{S} denotes the source image, and \mathbf{T} represents the target image. We assume \mathbf{F}_{ST} is the to-be-learned flow (*i.e.*, displacement field) that warps \mathbf{S} to \mathbf{T} . The optimization problem is typically defined as

$$\mathcal{F}(\mathbf{S}, \mathbf{T}) = \mathcal{L}_{similar}(\mathbf{T}, \Phi(\mathbf{S}, \mathbf{F}_{ST})) + \mathcal{R}(\mathbf{F}_{ST}), \quad (1)$$

where Φ denotes the transformation operator that warps the source image \mathbf{S} to the target image \mathbf{T} using the flow \mathbf{F}_{ST} . The first term in Eq. 1 is a similarity/matching/distance criterion, which is used to quantify the level of alignment between the warped source image $\Phi(\mathbf{S}, \mathbf{F}_{ST})$ and the target image \mathbf{T} . The second term is a regularizer that constrains the transformation to favor a specific property, such as encouraging the estimated flow to be locally smooth. The optimization problem consists of either maximizing or minimizing the objective function, depending on how the first term is defined.

Different algorithms for deformable registration mainly differ in deformable models, similarity criteria, and numerical optimization [1], [18]–[23]. In the literature, many types of similarity metrics have been proposed for image registration, such as mean squared distance (MSD) [24], [25], sum-of-squares distance (SSD) [26]–[28], normalized cross correlation (CC) [6], [11], [29], [30], and normalized mutual information (MI) [31]–[33]. Besides, there are various regularization terms developed to penalize undesired deformations, such as topology preservation [34]–[36], [36], [37], volume preservation [38]–[43], and rigidity constraints [18], [44], [45]. In general, the existing registration algorithms can be roughly divided into two categories, including 1) non-learning-based methods, and 2) learning-based methods. We now introduce relevant studies in these two categories.

2.2 Non-learning-based Registration Methods

Non-learning-based registration algorithms typically optimize a transformation iteratively based on an energy function in the form of Eq. 1. Based on how to compute the similarity between the warped source image and the target image, there are two types of registration approaches, including 1) volume-based methods where the voxel intensities in the whole volume are used to drive the registration process, and 2) landmark-based methods where features extracted at anatomical landmarks are employed to guide the matching of the local correspondence during registration.

The most popular non-learning-based methods for deformable registration include automatic image registration (AIR) [24], automatic registration toolbox (ART) [29], HAMMER [6], Demons [26], diffeomorphic Demons [27], [46], statistical parametric mapping (SPM) [25], deformable registration via attribute matching and mutual-saliency weighting (DRAMMS) [47], DROP [48], CC/MI/SSD-FFD [8], FNIRT [28], and symmetric normalization (SyN) [5]. Most of these methods require iterative optimization algorithms for parameter tuning [1]. Also, the registration performance of these methods would degrade when the source and target images have large variations in anatomical appearance. Therefore, robust and tuning-free deformable registration methods are highly desired for dealing with different data and registration tasks.

2.3 Learning-based Registration Methods

Many learning-based methods have been developed for deformable image registration [49], such as those based on random forest [50], support vector regression [51], sparse representation [52], and deep neural networks [12], [16], [17]. In these methods, deformable registration is often

formulated as a learning problem to estimate the registration parameters. Compared with non-learning methods, learning-based approaches can predict the transformation efficiently for unseen testing images based on pre-trained models. According to whether supervision information is needed, existing learning-based methods for image registration can be further categorized into two types: 1) supervised learning methods, and 2) unsupervised learning methods.

2.3.1 Supervised Methods

In supervised methods for deformable image registration, task-specific supervision information (*e.g.*, ground-truth registration) is usually required for model training. For instance, random forest has been applied for infant brain MRI registration and multi-modal image registration [50], [53], [54], based on hand-engineered imaging features and pre-defined ground-truth registration. However, the registration performance of such traditional learning-based methods could be sub-optimal, since the process of feature extraction is independent of the model training.

Several deep-learning-based methods have been recently developed, by incorporating feature extraction and registration model learning into a unified framework. Cao *et al.* [12] proposed a convolutional neural network (CNN) based regression model to directly learn the mapping from the input image pair (*i.e.*, target and source images) to their corresponding deformation fields. Yang *et al.* [55] developed a fully convolutional network (FCN) to predict 3D deformable registration, followed by a correction network to further refine the predicted transformation. Rohé [56] proposed an FCN model to learn the stationary velocity field, which consists of a contracting path to extract task-relevant features and a symmetric expanding path to output the transformation parameters. Sokooti *et al.* [57] proposed to predict displacement vectors by CNN models. Krebs *et al.* [58] adopted a deep reinforcement learning framework to estimate deformation fields. These methods usually require *task-specific ground-truth registration* for model learning. Since ground-truth information is difficult to collect, supervised methods usually have limited utility in practice. Also, the performance of supervised methods is largely determined by the quality of predefined ground-truth registration.

2.3.2 Unsupervised Methods

By maintaining the unsupervised nature of deformable registration, unsupervised deep-learning methods have also been applied for medical image registration [13]–[17]. That is, these methods do not rely on any pre-defined supervision information for network training. For example, Miao *et al.* [14] developed a CNN model to learn transformation parameters for 2D/3D images, where imaging features for parameter regression should be predefined, which means this network cannot be trained in an end-to-end manner. Wu *et al.* [13] proposed an unsupervised deep-learning algorithm for image registration based on image patches. Although this method can automatically extract features from images, it requires additional post-processing that cannot be handled inside CNNs. Shan *et al.* [15] developed an unsupervised end-to-end deep-learning model for 2D tissue registration, by directly predicting deformation field via a CNN. In [16], an end-to-end unsupervised deep-learning

model, consisting of a CNN-based regressor, a spatial transformer, and a re-sampler, was developed for deformable registration. Then, Guha *et al.* [17] proposed an unsupervised CNN model, in which a spatial transformer network (STN) is also used to reconstruct one image from another while imposing smoothness constraints on the registration field. This method has achieved superior accuracy for 3D image registration in comparison to previous methods.

It is worth noting that most of the existing deep-learning methods ignore the inherent inverse-consistent property of transformations between a pair of images [15]–[17]. That is, by independently estimating the transformation from an image **A** to an image **B** and that from **B** to **A**, these methods are unable to ensure that these transforms are inverse mappings of one another. Note that several studies tackle this shortcoming by jointly estimating the transformations from both **A** to **B** and **B** to **A**, under the consistency constraint that these transformations are inverses of one another [34], [59], [60]. However, these methods are not learning-based methods and require human-engineered feature representations (*e.g.*, image intensity with Fourier series parameterization) of input images, which may not be extracted in a task-oriented manner. Motivated by these studies, we develop an unsupervised deep network with an inverse-consistent constraint to encourage the inverse-consistent property of transformations between a pair of input images. Besides, considering using the smoothness constraint alone (as previous studies did) is not sufficient to guarantee that there is no folding in the estimated transformations, we further develop an anti-folding constraint to avoid foldings in the learned transformations.

3 PROPOSED METHOD

In this section, we first introduce the notations used in this paper. Then, we describe the proposed inverse-consistent deep neural network, as well as the objective function (with both the proposed inverse-consistent and anti-folding constraints) for network training. We finally introduce the implementation details for the proposed method.

3.1 Notations

In image registration, a pair of images are usually referred to as the source image and the target image. Since we do not rely on particular target images in our method, in this paper, we denote one input image as **A** and the other as **B**. These two images are defined in the image domain Ω . The transformation is a mapping function of the image domain Ω to itself, which spatially deforms any point locations to other locations. Also, we assume that the image **A** is deformed to match the image **B** according to a dense flow (*i.e.*, discrete displacement field) \mathbf{F}_{AB} defined in the **A** space, while the image **B** is deformed to match the image **A** via another dense flow \mathbf{F}_{BA} defined in the **B** space. Note that each element in a flow is a 3-dimensional vector (corresponding to three axes, *i.e.*, x, y, z), indicating the displacement of a particular voxel from its original location to a new location. In addition, the deformed/warped images of **A** and **B** are denoted as $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$, respectively.

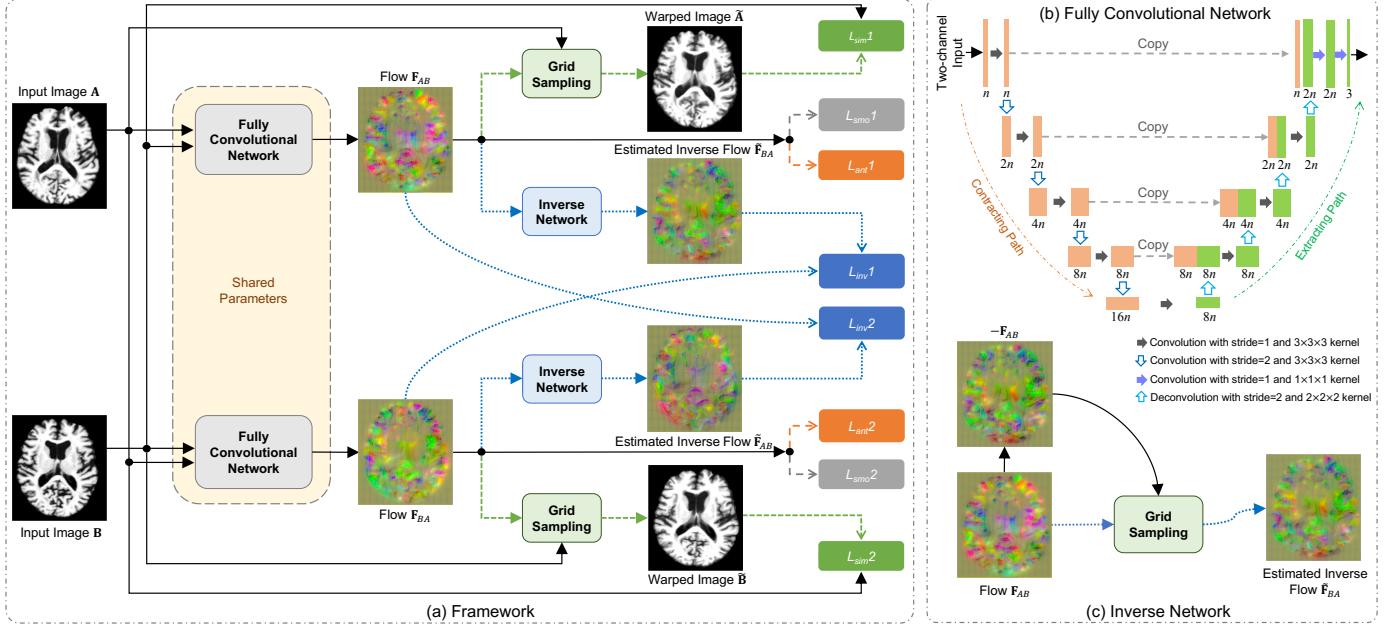


Fig. 2. Pipeline of the proposed Inverse-Consistent deep neural Network (ICNet) for unsupervised deformable image registration, which takes a pair of images as input. (a) Framework of ICNet, (b) architecture of fully convolutional network (FCN), and (c) illustration of the inverse network. The term n in (b) denotes the number of starting convolutional channels in FCN.

3.2 Inverse-consistent Unsupervised Neural Network

Figure 2 illustrates the proposed unsupervised deep network for deformable image registration. As shown in Fig. 2 (a), we employ a fully convolutional network (FCN) to model two dense, non-linear transformations (*i.e.*, \mathbf{F}_{AB} and \mathbf{F}_{BA}) from a pair of input images (*i.e.*, \mathbf{A} and \mathbf{B}) to their warped images (*i.e.*, $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$). There are two FCN modules in our proposed network. The first one is used to align the image \mathbf{A} to \mathbf{B} (as the target image) using the flow \mathbf{F}_{AB} , generating the warped image $\tilde{\mathbf{A}}$. In contrast, the second FCN is designed to model the registration mapping from the image \mathbf{B} to \mathbf{A} (as the target image) via the flow \mathbf{F}_{BA} , yielding the warped image $\tilde{\mathbf{B}}$. It is worth noting that these two FCNs share network structure and parameters.

As shown in Fig. 2 (b), the FCN we used here follows a U-Net architecture [61] to capture and combine both global and local structural information of input images. Specifically, the input data contain two channels (with each channel corresponds to a particular input image), and n is the number of starting filter channels of the FCN. The FCN contains a contracting path for image down-sampling and an expanding path for image up-sampling. Every step in the contracting path contains a $3 \times 3 \times 3$ convolution with a stride of 1, and a $3 \times 3 \times 3$ convolution with a stride of 2 for down-sampling. Besides, each step in the expanding path consists of a $2 \times 2 \times 2$ deconvolution with a stride of 2 for up-sampling, followed by a concatenation process to combine up-sampled feature maps with the corresponding feature maps from the contracting path, and then a $3 \times 3 \times 3$ convolution with a stride of 1. In this network, each convolution is followed by a rectified linear unit (ReLU) activation, while the output of the last layer (having 3 filter channels that are corresponding to the x , y , and z axis) is constrained into $[-\tau, \tau]$. That is, we first use a \tanh function to normalize the output of the last layer to $[-1, 1]$, followed by multiplying by a

constant τ (*i.e.*, the maximum displacement magnitude). Since both input images are treated equally in the proposed ICNet framework, the two FCNs in Fig. 2 (b) share the same parameters.

Besides, we can see from Fig. 2 (a) that a grid sampling module is utilized to generate the warped image (*e.g.*, $\tilde{\mathbf{A}}$), based on the input image (*e.g.*, \mathbf{A}) and the learned flow (*e.g.*, \mathbf{F}_{AB}). Specifically, such grid sampling is implemented via the fully-differentiable spatial transformer network (STN) [62], containing a regular spatial grid generator and a sampler. The flow (displacement field) predicted by our image registration network is used to transform the regular spatial grid into a sampling grid. Then, the sampler uses the sampling grid to warp the input image. Bilinear interpolation is used during the sampling process, making STN fully differentiable for back propagation.

Furthermore, an inverse network is developed to generate an estimated inverse flow (*e.g.*, $\tilde{\mathbf{F}}_{BA}$) of each transformation learned by the FCN (*e.g.*, \mathbf{F}_{AB}), based on which an inverse-consistent loss (*i.e.*, $Loss_{inv1}$ and $Loss_{inv2}$) is further adopted to encourage the inverse-consistent property of two transformations. As shown in Fig. 2 (c), we utilize the grid sampling strategy to generate the estimated inverse flow $\tilde{\mathbf{F}}_{BA}$, based on both \mathbf{F}_{AB} and $-\mathbf{F}_{AB}$.

3.3 Proposed Objective Function

3.3.1 Inverse-consistent Constraint

Existing deep-learning methods typically ignore the inverse-consistent property of transformations between a pair of images [15]–[17]. Motivated by previous non-learning-based inverse-consistent methods [34], [59], [60], we propose to simultaneously estimate the transformation from \mathbf{A} to \mathbf{B} (*i.e.*, \mathbf{F}_{AB}) and the transformation from \mathbf{B} to \mathbf{A} (*i.e.*, \mathbf{F}_{BA}), and enforce the consistency constraint that these bidirectional transformations are inverse mappings of one another.

Specifically, we propose an *inverse-consistent regularization* term to penalize the difference between two transformations from the respective inverse mappings. As shown in Fig. 2 (c), we rely on an inverse network to generate the inverse mapping (*e.g.*, $\tilde{\mathbf{F}}_{BA}$) of each transformation (*e.g.*, \mathbf{F}_{AB}). Specifically, for the flow \mathbf{F}_{AB} , we first obtain its negative flow (*i.e.*, $-\mathbf{F}_{AB}$) in the \mathbf{A} space. We then feed both \mathbf{F}_{AB} and $-\mathbf{F}_{AB}$ to the grid sampling module (via a STN) to obtain the estimated inverse flow (*i.e.*, $\tilde{\mathbf{F}}_{BA}$ in the \mathbf{B} space) of \mathbf{F}_{AB} . Similarly, we feed both \mathbf{F}_{BA} and its negative flow $-\mathbf{F}_{BA}$ to the grid sampling module, and hence can obtain the estimated inverse flow (*i.e.*, $\tilde{\mathbf{F}}_{AB}$) of \mathbf{F}_{BA} . Mathematically, the proposed inverse-consistent constraint can be defined as follows

$$\mathcal{L}_{inv} = \|\mathbf{F}_{AB} - \tilde{\mathbf{F}}_{AB}\|_F^2 + \|\mathbf{F}_{BA} - \tilde{\mathbf{F}}_{BA}\|_F^2 \quad (2)$$

with

$$\begin{aligned} \tilde{\mathbf{F}}_{AB} &= \mathcal{G}(\mathbf{F}_{BA}, -\mathbf{F}_{BA}) \\ \tilde{\mathbf{F}}_{BA} &= \mathcal{G}(\mathbf{F}_{AB}, -\mathbf{F}_{AB}) \end{aligned} \quad (3)$$

where \mathcal{G} is the mapping generated by the grid sampling module (via a STN), and $\|\cdot\|_F$ represents the Frobenius norm of a matrix. The two terms in Eq. 2 correspond to the notations L_{inv1} and L_{inv2} in Fig. 2 (a). By minimizing Eq. 2, we concurrently encourage both the difference between the flows \mathbf{F}_{AB} and $\tilde{\mathbf{F}}_{AB}$ (*i.e.*, the inverse of \mathbf{F}_{BA}) and that between \mathbf{F}_{BA} and $\tilde{\mathbf{F}}_{BA}$ (*i.e.*, the inverse of \mathbf{F}_{AB}) to be small. In this way, the inverse-consistent property of the to-be-estimated transformations can be explicitly modeled in the proposed network.

3.3.2 Anti-folding Constraint

As mentioned in Section 1 (*e.g.*, Fig. 1), if we excessively encourage local smoothness of the to-be-estimated flow by using a large weight for the smoothness constraint, the registration results will be inaccurate. Otherwise, there will be possible foldings in the flow, thus yielding unreasonable registration. To deal with this issue, besides using the conventional smoothness constraint, we also develop an anti-folding constraint as

$$\begin{aligned} \mathcal{L}_{ant} = \sum_{p \in \Omega} \sum_{i \in \{x, y, z\}} \delta(\nabla \mathbf{F}_{AB}^i(p) + 1) |\nabla \mathbf{F}_{AB}^i(p)|^2 \\ + \delta(\nabla \mathbf{F}_{BA}^i(p) + 1) |\nabla \mathbf{F}_{BA}^i(p)|^2 \end{aligned} \quad (4)$$

where $\nabla \mathbf{F}_{AB}^i(p)$ is the gradient of the flow \mathbf{F}_{AB} along the i -th ($i \in \{x, y, z\}$) axis at the location of the voxel p . Besides, the term $\delta(Q)$ is an index function used to penalize the gradient of the flow at the locations with foldings. That is, if $Q \leq 0$, $\delta(Q) = |Q|$; and $\delta(Q) = 0$, otherwise.

The purpose of Eq. 4 can be explained as follows. If there is a folding at the location of p along the i -th axis (*i.e.*, $\nabla \mathbf{F}_{AB}^i(p) + 1 \leq 0$), we enforce the penalty $|\nabla \mathbf{F}_{AB}^i(p) + 1|$ on the gradient at this location, requiring this gradient to be small. In contrast, if $\nabla \mathbf{F}_{AB}^i(p) + 1 > 0$ (*i.e.*, no folding at the location of p along the i -th axis), we do not penalize the gradient at this location. More detailed explanation can be found in the Appendix.

3.3.3 Smoothness Constraint

In previous studies, the to-be-estimated deformation field is generally to be locally smoothed via a smoothness constraint on its spatial gradients [16], [17]. Here, we also use such smoothness constraint in the objective function as

$$\mathcal{L}_{smo} = \sum_{p \in \Omega} (\|\nabla \mathbf{F}_{AB}(p)\|_2^2 + \|\nabla \mathbf{F}_{BA}(p)\|_2^2) \quad (5)$$

where $\nabla \mathbf{F}_{AB}(p)$ is the gradient of the flow \mathbf{F}_{AB} at the voxel p , while $\nabla \mathbf{F}_{BA}(p)$ denotes the gradient of the flow \mathbf{F}_{BA} at the voxel p . The operation $\|\cdot\|_2$ represents the l_2 norm of a vector. Here, we approximate the spatial gradients using the differences between neighboring voxels.

3.3.4 Objective Function

In this work, we utilize the mean squared distance (MSD) as the similarity metric to compare the alignment between the warped image and its corresponding target image. Specifically, the MSD-based symmetric similarity is employed to measure the shape differences between the deformed image $\tilde{\mathbf{A}}$ and image \mathbf{B} and the differences between the deformed image $\tilde{\mathbf{B}}$ and image \mathbf{A} , which is defined as

$$\mathcal{L}_{sim} = \|\mathbf{B} - \tilde{\mathbf{A}}\|_F^2 + \|\mathbf{A} - \tilde{\mathbf{B}}\|_F^2 \quad (6)$$

where $\tilde{\mathbf{A}} = \mathcal{G}(\mathbf{F}_{AB}, \mathbf{A})$ and $\tilde{\mathbf{B}} = \mathcal{G}(\mathbf{F}_{BA}, \mathbf{B})$, and \mathcal{G} is the mapping function learned in the grid sampling module.

Accordingly, the loss function of our proposed ICNet for deformable registration is formulated as follows

$$\begin{aligned} \mathcal{L}(\mathbf{A}, \mathbf{B}) = & \mathcal{L}_{sim} + \alpha \mathcal{L}_{smo} \\ & + \beta \mathcal{L}_{inv} + \gamma \mathcal{L}_{ant} \end{aligned} \quad (7)$$

where the parameters α , β and γ are used to balance the contributions of the smoothness, inverse-consistent and anti-folding regularizers, respectively.

3.4 Implementation

We implement the proposed deep network with Pytorch [63]. The objective function in Eq. 7 is optimized by the Adam algorithm [64] combined with a back-propagation algorithm for computing gradients as well as updating network parameters. The learning rate for Adam is empirically set to 5×10^{-4} . In Fig. 3, we report the change curves of both training (red) and validation (green) losses on the public Alzheimer’s Disease Neuroimaging Initiative (ADNI-1) database [65], where 10% subjects are randomly selected as the validation data, and the remaining subjects are used as the training data. This figure indicates that the proposed ICNet method generalizes well with almost no over-fitting issue, and the proposed objective functions converges quickly within 40,000 iterations. For readers’ convenience, our code and trained model will be made publicly available online.

4 EXPERIMENTS

In this section, we first introduce the studied materials, competing methods, and experimental settings. Then, we present results of brain tissue segmentation and anatomical landmark detection based on the warped MR images achieved by different registration methods. We finally analyze the computational costs of different methods.

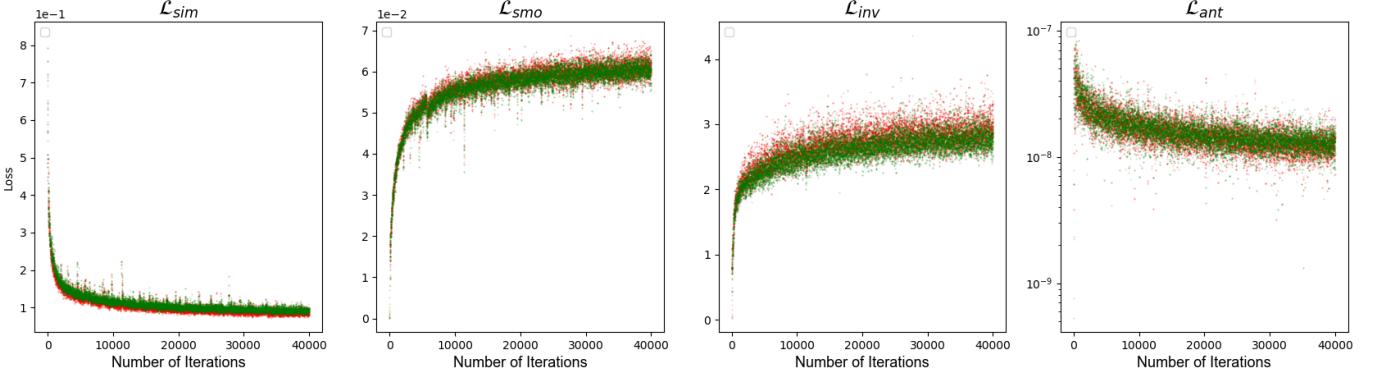


Fig. 3. Training and validation loss of the proposed ICNet regarding (a) similarity term, (b) inverse-constraint regularizer, (c) anti-folding regularizer, and (d) smooth regularizer. Red lines denote the training loss, while green lines correspond to the validation loss.

4.1 Materials and Image Pre-processing

We perform experiments on 860 subjects from two subsets of the public ADNI database¹ [65], *i.e.*, 1) ADNI-1, and 2) ADNI-2. To be specific, there are 805 subjects with the baseline structural brain MRI scans in ADNI-1, while the remaining 55 subjects with the baseline structural MRI data are randomly selected from ADNI-2. Since several subjects participated in both ADNI-1 and ADNI-2, we simply remove these subjects from ADNI-2, to ensure that ADNI-1 and ADNI-2 are independent datasets in the experiments. Notably, the studied subjects from ADNI-1 have 1.5 T T1-weighted MRI scans, while those in ADNI-2 have 3.0 T T1-weighted MRI data.

For all structural brain MR images, we perform both *spatial normalization* and *intensity normalization* for image normalization. For spatial normalization, we first perform skull stripping and cerebellar removal for all brain MRIs, and then linearly align them to a common Colin27 [66] template. We further resample all linearly aligned images to have the same spatial resolution (*i.e.*, 1 mm × 1 mm × 1 mm), followed by cropping them to have the same image size (*i.e.*, 144 × 192 × 160). For intensity normalization, we first match the intensity histogram of each brain MRI to that of the Colin27 template by using a histogram matching algorithm. Then, we also perform the z-score normalization to make the mean intensity of each image is zero and the standard deviation is one.

In the experiments, we perform two tasks to evaluate the registration performance of different methods, including 1) *brain tissue segmentation* and 2) *anatomical landmark detection*. In the task of brain tissue segmentation, the ground-truth segmentation is generated by using first FAST in FSL [67] to obtain the tissue segmentation map, followed by further manual correction. As illustrated in Fig. 4 (b), three tissues are segmented from each brain MR image, including cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM). In the task of anatomical landmark detection, the ground-truth landmarks are manually annotated by an experienced radiologist. As shown in Fig. 4 (c), five anatomical landmarks are annotated in each brain MR image, which are mainly located in ventricles.

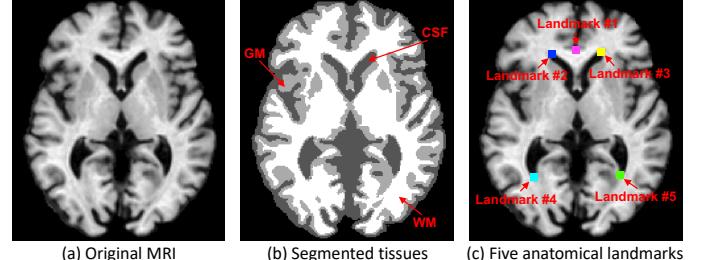


Fig. 4. Illustration of ground-truth tissue segmentations and anatomical landmarks: (a) the original MRI, (b) three types of segmented tissues (*i.e.*, CSF, GM, and WM), and (c) five anatomical landmarks denoted by blocks with different colors.

4.2 Competing Methods

We compare the proposed ICNet method with three state-of-the-art methods for deformable image registration, including 1) Demons with the symmetric local correlation coefficient used as the similarity metric [46], 2) symmetric normalization (denoted as SyN) [5], and 3) a unsupervised deep-learning (denoted as DL) method with a minimum mean squared error (MMSE) loss and a smoothness constraint [17]. Among them, Demons and SyN are non-learning-based methods, while DL is an unsupervised learning method. For the fair comparison, the network architecture of the DL method is similar to our ICNet, while its objective function is different from ours. Specifically, only the first two terms in Eq. 7 is included in the objective function of DL, and hence, DL can be regarded as a degenerated variant of ICNet.

To evaluate the effectiveness of the proposed two constraints (*i.e.*, the inverse-consistent constraint and the anti-folding constraint), we further compare ICNet with its two variants. The first variant is denoted as **ICNet-1**, in which the proposed inverse-consistent constraint is removed. Similarly, the second variant is denoted as **ICNet-2**, in which the proposed anti-folding constraint is not used.

4.3 Experimental Settings

Since the six methods for comparison are all unsupervised, we do not need to generate any ground-truth registration for them. For the non-learning-based methods (*i.e.*, Demons and SyN), we utilize their recommended parameter settings in the experiments. For the learning-based methods (*i.e.*,

1. <http://adni.loni.usc.edu>

TABLE 1
Segmentation results for three brain tissues (*i.e.*, CSF, GM and WM), achieved by six different methods on the ADNI dataset.

Tissue	Metric	Demons	SyN	DL	ICNet-1	ICNet-2	ICNet
CSF	DSC (%)	77.16 ± 1.82	78.43 ± 1.38	79.76 ± 1.22	80.75 ± 1.31	82.08 ± 1.14	83.58 ± 1.17
	SEN (%)	71.99 ± 2.15	76.57 ± 1.96	77.38 ± 1.94	79.00 ± 1.84	80.16 ± 1.75	82.55 ± 1.80
	PPV (%)	83.16 ± 2.07	80.41 ± 1.47	82.34 ± 1.38.	82.61 ± 1.44	84.13 ± 1.36	84.66 ± 1.31
	ASD	0.69 ± 0.05	0.73 ± 0.03	0.68 ± 0.03	0.65 ± 0.03	0.60 ± 0.03	0.56 ± 0.03
	HD	11.63 ± 0.84	11.74 ± 0.78	11.60 ± 0.80	11.41 ± 0.74	11.29 ± 0.71	11.10 ± 0.78
GM	DSC (%)	76.15 ± 1.33	74.16 ± 1.10	75.04 ± 1.06	77.93 ± 1.05	78.30 ± 1.04	80.59 ± 0.98
	SEN (%)	74.71 ± 1.63	74.23 ± 1.20	74.41 ± 1.22	77.96 ± 1.24	77.92 ± 1.25	82.28 ± 1.15
	PPV (%)	77.65 ± 1.38	74.11 ± 1.44	75.70 ± 1.41	77.92 ± 1.38	78.70 ± 1.37	78.99 ± 1.38
	ASD	0.58 ± 0.03	0.65 ± 0.02	0.63 ± 0.02	0.58 ± 0.02	0.56 ± 0.02	0.54 ± 0.02
	HD	9.85 ± 1.25	9.13 ± 0.92	9.13 ± 1.01	9.20 ± 1.04	8.95 ± 0.95	9.07 ± 1.05
WM	DSC (%)	84.02 ± 1.75	84.67 ± 1.10	86.03 ± 1.01	86.92 ± 1.11	87.82 ± 1.00	88.13 ± 0.94
	SEN (%)	80.99 ± 2.04	82.79 ± 1.75	85.35 ± 1.63	85.03 ± 1.73	86.17 ± 1.71	86.56 ± 1.61
	PPV (%)	87.31 ± 2.05	86.66 ± 1.25	86.74 ± 1.24	88.92 ± 1.32	89.13 ± 1.24	90.22 ± 1.19
	ASD	0.80 ± 0.06	0.88 ± 0.04	0.81 ± 0.04	0.76 ± 0.04	0.71 ± 0.04	0.71 ± 0.04
	HD	13.73 ± 1.38	13.40 ± 1.13	12.77 ± 1.22	13.18 ± 1.28	13.16 ± 1.24	12.71 ± 1.11

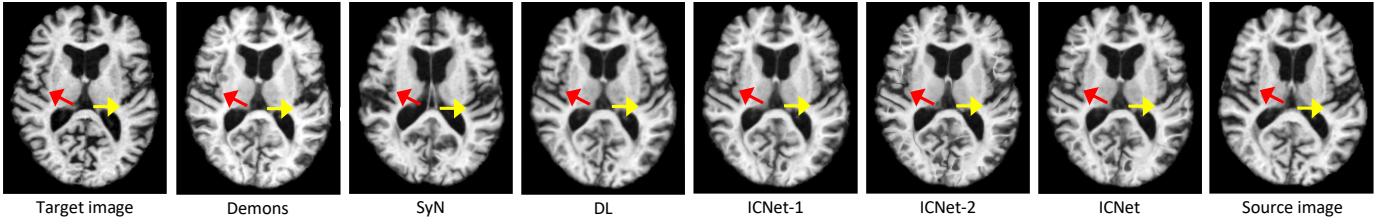


Fig. 5. Registration results achieved by six different methods for deforming a source image (right) to a target image (left). Red and yellow arrows indicate the right and left planum temporale, respectively.

DL, ICNet-1, ICNet-2, and ICNet), we treat the ADNI-1 and ADNI-2 as the training set and testing set, respectively. We randomly select 10% subjects from the ADNI-1 as the validation data, while the remaining subjects are used as the training data.

In the *training* stage, we randomly select a pair of MR images from the ADNI-1 as the input for each learning-based method. In the *testing* stage for all competing methods, to perform deformable registration, we first select 5 MR images from the ADNI-2 as atlas images, while the remaining 50 MR images are used as testing images. By using different deformable registration algorithms, we first warp each of the 5 atlases to a particular testing image, and thus generating 5 warped atlas images based on this testing image. Then, we employ a multi-atlas based segmentation algorithm with a majority voting strategy [68] to perform brain tissue segmentation in each testing image. Similarly, for landmark detection, each landmark in the atlases is first mapped onto a particular testing image via the corresponding deformable transformation [69]. Thus, given a testing image, we obtain 5 warped landmark positions for each landmark, followed by averaging these positions to generate the final location of this landmark. It is worth noting that, to evaluate the performance of the proposed ICNet for deformable image registration, we only utilize atlas-based methods for tissue segmentation and landmark detection, while other supervised learning methods are beyond the scope of this paper.

In ICNet, the parameter γ is empirically set to 10^5 to avoid folding in the flow as much as possible, and the other two parameters (*i.e.*, α and β) are determined via grid search within the range of $[10^{-5}, 10^{-4}, \dots, 10^5]$ on the

validation set. Similarly, parameters in the three competing methods (*i.e.*, α in DL, α in ICNet-1, and α, β in ICNet-2) are also selected from the same range through cross-validation. Besides, we set $\beta = 0, \gamma = 10^5$ in ICNet-1, and $\gamma = 0$ in ICNet-2. The number of starting filter channels n (see Fig. 2 (b)) for ICNet and its two variants (*i.e.*, ICNet-1 and ICNet-2) is empirically set to 8. In each of the four deep-learning methods (*i.e.*, DL, ICNet-1, ICNet-2, and ICNet), the output of the last layer (with 3 filter channels corresponding to the x , y , and z axis) is constrained into the range $[-\tau, \tau]$. Following [70], we empirically set $\tau = 7$ in this work, considering the displacement magnitude is usually less than 7 in the deformable registration of brain MRI scans.

In the experiments of brain tissue segmentation, five complementary metrics are used for quantitative evaluation of segmentation performance, including 1) dice similarity coefficient (DSC), 2) sensitivity (SEN), 3) positive predictive value (PPV), 4) average symmetric surface distance (ASD), and 5) Hausdorff distance (HD). In the experiments of anatomical landmark detection, for each landmark, we report the landmark detection error by computing the Euclidean distance between the estimated landmark location (achieved by a specific method) and its ground-truth location. For the evaluation metrics of ACC, SEN and PPV, higher values indicate better performance. For the remaining three metrics (*i.e.*, ASD, HD, and detection error), lower values denote better performance.

4.4 Results of Brain Tissue Segmentation

In the first group of experiments, we perform the segmentation of three types of brain tissues (*i.e.*, CSF, GM and WM),

TABLE 2
Landmark detection error (mm) achieved by six different methods on the ADNI dataset.

Index	Demons	SyN	DL	ICNet-1	ICNet-2	ICNet
Landmark #1	2.40 ± 0.96	2.44 ± 0.96	2.18 ± 0.87	2.19 ± 0.83	2.12 ± 0.91	2.10 ± 0.91
Landmark #2	3.07 ± 1.35	2.77 ± 1.26	2.65 ± 1.14	2.53 ± 1.07	2.66 ± 1.15	2.68 ± 1.16
Landmark #3	3.61 ± 1.84	3.56 ± 1.78	3.02 ± 1.69	3.03 ± 1.80	2.75 ± 1.63	2.60 ± 1.57
Landmark #4	3.41 ± 1.30	3.23 ± 1.43	2.78 ± 1.29	2.82 ± 1.36	2.72 ± 1.18	2.73 ± 1.16
Landmark #5	3.32 ± 1.21	3.41 ± 1.34	3.06 ± 1.19	3.06 ± 1.25	2.95 ± 1.15	2.94 ± 1.14
Average	3.16 ± 1.42	3.08 ± 1.44	2.74 ± 1.31	2.73 ± 1.34	2.64 ± 1.26	2.61 ± 1.24

based on the warped atlas images generated by six different methods. The experimental results are shown in Table 1.

From Table 1, one could have the following observations. *First*, in most cases, the proposed methods (*i.e.*, ICNet-1, ICNet-2, and ICNet) achieve the overall best performance (regarding DSC, SEN, PPV, ASD, and HD) for segmenting all the three types of tissues. For instance, the DSC value achieved by ICNet for CSF segmentation is 83.58%, while the DSC produced by the conventional deep-learning method (*i.e.*, DL) is only 79.76%. *Second*, even though no supervision information is required, four unsupervised deep-learning-based methods (*i.e.*, DL, ICNet-1, ICNet-2, and ICNet) usually outperform two non-learning-based methods (*i.e.*, Demons and SyN). The underlying reason may be that deep-learning-based methods can extract task-oriented features via neural networks, while conventional methods simply employ hand-engineered features of brain MRIs. *Besides*, we can see that our ICNet method usually outperform its two variants (*i.e.*, ICNet-1 and ICNet-2). Note that ICNet-1 does not use the proposed inverse-consistent constraint, and ICNet-2 does not utilize the proposed anti-folding constraint. This implies that including both the inverse-consistent and anti-folding constraints to train our ICNet could boost the deformable image registration performance.

Given a testing image, we further visually compare the registration results for a source brain MR image achieved by different methods in Fig. 5. From Fig. 5, we can see that the proposed ICNet method brings impressive improvement for the registration results, compared with the competing methods. For instance, it is obvious that the regions of the left and right planum temporale are more accurately registered to the target image using ICNet, as indicated by the red arrow (left planum temporale) and the yellow arrow (right planum temporale) in Fig. 5.

4.5 Results of Anatomical Landmark Detection

In the second group of experiments, we perform landmark detection based on the deformed atlas images generated by different registration methods, with the results reported in Table 2. From Table 2, we can see that the overall performance of our ICNet method is superior to the five competing methods. For instance, the average landmark detection error achieved by ICNet is 2.61, which is lower than the result of Demons (*i.e.*, 3.16).

It is worth noting that our methods (*i.e.*, ICNet-1, ICNet-2, and ICNet) are unsupervised and do not need to generate the ground-truth flow for each to-be-registered image during network training. This is a particularly useful property for the deformable registration algorithm, which not only

maintains the unsupervised nature of deformable registration, but also avoids the challenge of collecting accurate ground-real registration.

4.6 Computational Cost

We now analyze the computational costs of the proposed ICNet method and those competing methods. For the four deep-learning-based methods (*i.e.*, DL, ICNet-1, ICNet-2, and ICNet), the training process is performed off-line, while the non-learning-based methods (*i.e.*, Demons and SyN) do not need any training process. Hence, we only analyze the on-line computational cost for nonlinearly registering a new brain MRI in the application/testing stage. Table 3 reports the computational costs of different methods. Note that Demons and SyN are implemented using a CPU (i7-7700, 3.6GHz), while the remaining four methods are implemented using a GPU (GTX 1080ti). We can observe from Table 3 that the computational costs of the four deep-learning-based methods require only ~ 0.25 second for deformable registration of one MRI, which is much faster than Demons (~ 1 minutes) and SyN (~ 2 hours). These results further demonstrate the potential utility of our method in practical applications.

TABLE 3
Computation time of different methods for deformable registration of each testing image in the application stage.

Method	Demons	SyN	DL	ICNet-1	ICNet-2	ICNet
Time	$\sim 1m$	$\sim 2h$	$\sim 0.25s$	$\sim 0.25s$	$\sim 0.25s$	$\sim 0.25s$

5 DISCUSSION

In this section, we first investigate the effect of two essential components (*i.e.*, the inverse-consistent and anti-folding constraints) in the proposed ICNet method. We then analyze the influence of network architectures and a network refining strategy used in the application stage.

5.1 Influence of Inverse-consistent Constraint

To evaluate the influence of the proposed inverse-consistent constraint in Eq. 2, we visually illustrate the flows estimated by ICNet with different contributions from the proposed inverse-consistent constraint. Fig. 6 shows a pair of input images, as well as the flows and estimated inverse flows achieved by ICNet-1 and ICNet with different parameter settings. Here, we fix the parameter $\gamma = 10^5$ for the anti-folding constraint, while the inverse flows are generated by linear interpolation via the proposed inverse network (see

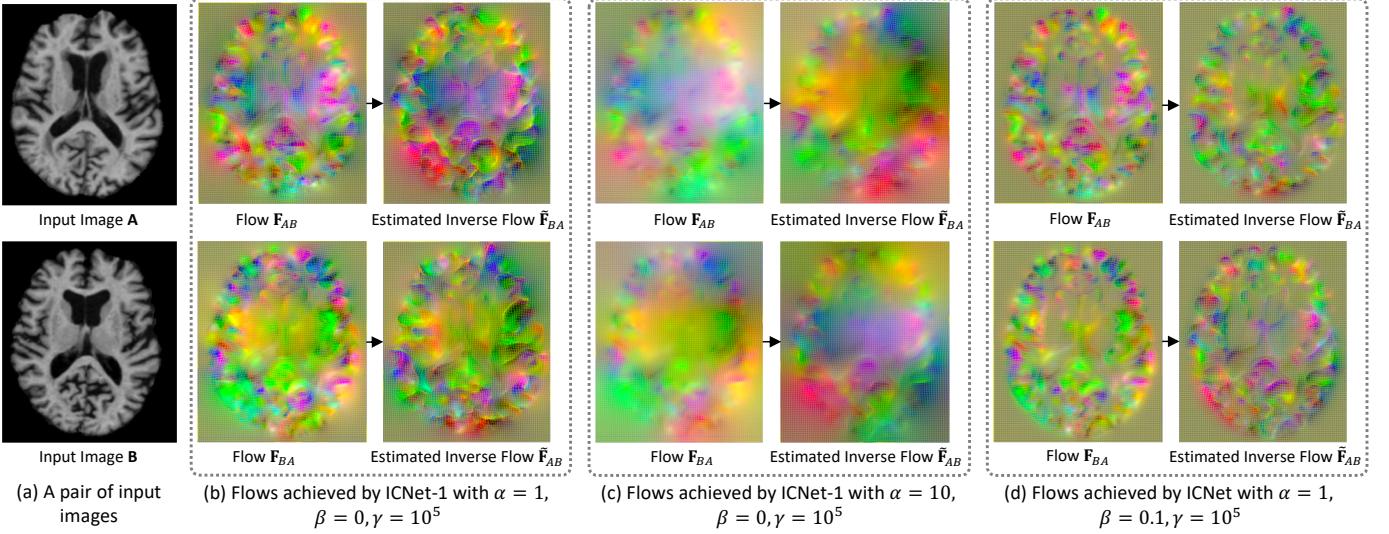


Fig. 6. Effect of the inverse-consistent constraint. (a) A pair of input images. (b) Flows and estimated inverse flows generated by ICNet-1 with a small weight for the smoothness constraint and without the inverse-consistent constraint (*i.e.*, $\alpha = 1$, $\beta = 0$ and $\gamma = 10^5$). (c) Flows and estimated inverse flows yielded by ICNet-1 with a large weight for the smoothness constraint and without the inverse-consistent constraint (*i.e.*, $\alpha = 10$, $\beta = 0$ and $\gamma = 10^5$). (d) Flows and estimated inverse flows generated by ICNet with both the smoothness and inverse-consistent constraints (*i.e.*, $\alpha = 1$, $\beta = 0.1$ and $\gamma = 10^5$).

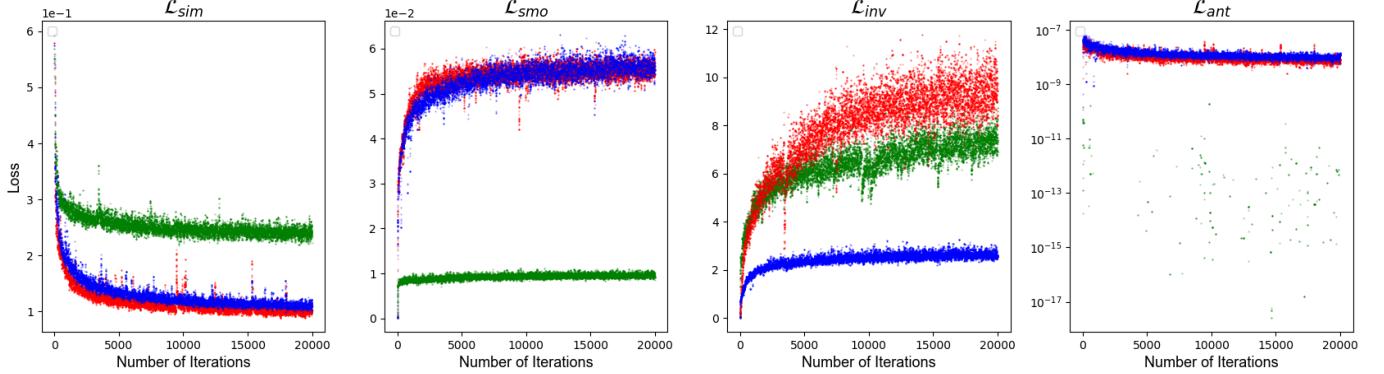


Fig. 7. Validation losses achieved by ICNet-1 and ICNet using different weights for the smoothing and inverse-consistent constraints, regarding four terms (from left to right) in Eq. 7. Red line denotes the validation loss of ICNet-1 using a small weight for the smoothness constraint and without the inverse-consistent constraint (*i.e.*, $\alpha = 1$, $\beta = 0$, and $\gamma = 10^5$), green line represents the validation loss of ICNet-1 with a large weight for the smoothness constraint and without the inverse-consistent constraint (*i.e.*, $\alpha = 10$, $\beta = 0$, and $\gamma = 10^5$), and blue line denotes the validation loss of ICNet using a small weight for the smoothness constraint and the inverse-consistent constraint (*i.e.*, $\alpha = 1$, $\beta = 0$, and $\gamma = 10^5$).

Fig. 2 (c)). Results in Fig. 6 (b)-(c) are generated by ICNet-1 without using the inverse-consistent constraint (*i.e.*, $\beta = 0$) but having different weights for the smoothness constraint, while those in Fig. 6 (d) are yielded by ICNet with the inverse-consistent constraint.

From Fig. 6 (b), we can observe that using a small weight (*i.e.*, $\alpha = 1$) for the smoothness term in ICNet-1 cannot generate good results (with foldings in the estimated inverse flow), and also the flow between two images is not inverse-consistent. For instance, the estimated inverse flow $\tilde{\mathbf{F}}_{AB}$ is not consistent with \mathbf{F}_{AB} , while $\tilde{\mathbf{F}}_{BA}$ looks different from \mathbf{F}_{BA} . Fig. 6 (c) suggests that using a large weight (*i.e.*, $\alpha = 10$) for the smoothness constraint in ICNet-1 will generate over smooth flow, which may degrade the registration accuracy. Fig. 6 (d) shows that ICNet ($\alpha = 1$, $\beta = 0.1$) can generate flows with a reasonable smoothness degree. Also, it can be seen from Fig. 6 (d) that $\tilde{\mathbf{F}}_{AB}$ is similar to \mathbf{F}_{AB} , and $\tilde{\mathbf{F}}_{BA}$ also looks similar to \mathbf{F}_{BA} . This suggests that ICNet can well preserve the inverse-consistent property of

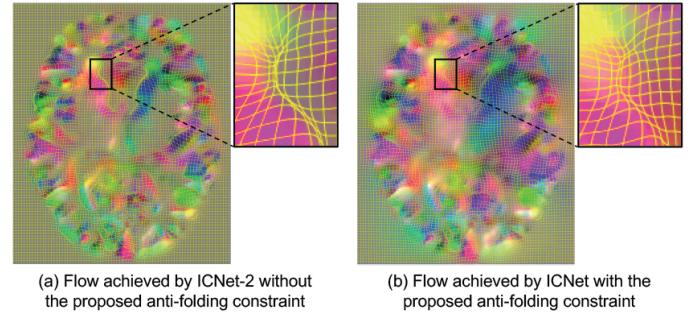


Fig. 8. Effect of using the anti-folding constraint, with flows generated by (a) ICNet-2 without the anti-folding constraint (*i.e.*, $\gamma = 0$) and (b) ICNet with the anti-folding constraint (*i.e.*, $\gamma = 10^5$). Here, $\alpha = 1$ and $\beta = 0.1$.

the bidirectional flows.

Furthermore, we show the validation loss achieved by ICNet-1 and ICNet with different contributions from the smoothing and inverse-consistent constraints in Fig. 7. For ICNet-1 with $\beta = 0$, the inverse-consistent term in Eq. 7 is

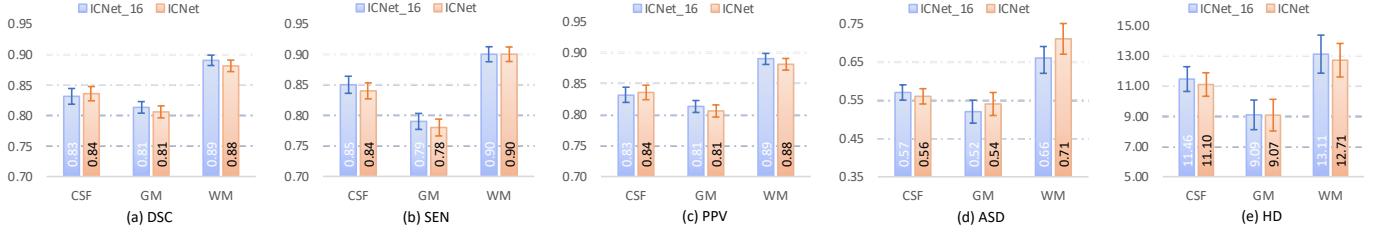


Fig. 9. Effect of using different architectures for the proposed method. Here, ICNet has $n = 8$ starting filter channels, while ICNet_16 (a variant of ICNet) includes $n = 16$ starting filter channels. (a)-(e) show the tissue segmentation results in terms of five evaluation metrics.

not used for network optimization and we only record the corresponding loss here. In Fig. 7, red line denotes the validation loss of ICNet-1 with $\alpha = 1$, green line represents the loss of ICNet-1 with $\alpha = 10$ for the smoothness constraint, and blue line denotes the loss of ICNet with $\alpha = 1$ and $\beta = 0.1$. As shown in the figure, using a large weight for the smoothing term (green lines) can have relatively small loss \mathcal{L}_{smo} , but the \mathcal{L}_{inv} is pretty large. It implies the warped source image is largely different from the target image. Besides, using a small weight for the smoothness regularizer (red lines) can yield relatively large loss \mathcal{L}_{inv} , but a good loss \mathcal{L}_{sim} concerning the similarity between the warped source image and the target image. In contrast, the losses of ICNet with $\alpha = 1$ and $\beta = 0.1$ (blue lines) suggest that ICNet can not only produce inverse-consistent registration, but also keep the warped source image as similar as possible to the target image.

5.2 Influence of Anti-folding Constraint

We then study the influence of the proposed anti-folding constraint, by comparing ICNet with ICNet-2 (without using the anti-fold constraint). In this group of experiments, we fix the parameters for the smoothing and inverse-consistent constraints (*i.e.*, $\alpha = 1$ and $\beta = 0.1$). Fig. 8 shows the flows generated by ICNet-2 (left) and ICNet (right). It can be observed from Fig. 8 (a) that the flow generated by ICNet-2 without using the anti-folding regularizer includes many folding (see the black rectangle) that would degrade the registration accuracy. In contrast, Fig. 8 (b) suggests that ICNet using the anti-folding regularizer can effectively avoid the folding in the flow. These results demonstrate the effectiveness of the anti-folding constraint in preventing foldings in the learned transformation.

5.3 Influence of Network Architecture

We also investigate the influence of the network architecture on the performance of ICNet, where the number of starting filter channels in FCN (see n in Fig. 2 (b)) is the essential component. In the above-mentioned experiments, we empirically set $n = 8$. In this group of experiments, we compare ICNet with its variant denoted as ICNet_16 with $n = 16$, and report the results of tissue segmentation achieved by these two methods in Fig. 9.

It can be seen from Fig. 9 that ICNet_16 achieves slightly better results in segmenting the three types of tissues, compared with ICNet. This implies that using more filter channels in the FCN within the proposed ICNet framework helps boost the registration accuracy, thus improving the tissue segmentation performance. Note that ICNet_16 requires

much large memory ($\sim 20GB$) for network training, while the ICNet with 8 starting filter channels only need $\sim 10GB$ memory. Also, the training time of ICNet_16 will double that of ICNet. Considering the marginal improvement of the registration performance shown in Fig. 9, we can flexibly choose the number of starting filter channels in practice, based on the memory capacity at hand.

5.4 Influence of Network Refining Strategy

The proposed ICNet is unsupervised, without using any ground-truth registration results. Therefore, given a pair of new testing images, we can feed them to ICNet (trained on the training data) to further refine the network, thus adapting the network to the testing images. Here, we denote ICNet with such a network refining process as ICNet_R. In ICNet_R, we first optimize the network parameters using the training data, and then refine the network (with the learned parameters as initialization) for the new pair of testing images. In the refining stage, we use a small learning rate (*i.e.*, 1×10^{-5}) for optimization, and the number of iteration is empirically set to 100. After refinement, we can use the newly learned network parameters to produce the final registration results for the testing images. The experimental results on tissue segmentation achieved by ICNet and its refined variant (*i.e.*, ICNet_R) are shown in Fig. 10.

It can be seen from Fig. 10 that ICNet_R consistently outperforms ICNet in segmenting three types of tissues, regarding five evaluation metrics. For instance, the PPV value of ICNet_R in segmenting CSF is 0.87, while that of ICNet is only 0.84. The possible reason could be that the refining strategy makes the network to be better coordinated with the new input images, thus reducing the negative influence of distribution differences between training and test data. It is worth noting that such network refining strategy is a general approach, which can also be applied to improving other unsupervised algorithms for image registration.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose an inverse-consistent deep network (ICNet) for unsupervised deformable image registration. Specifically, we develop an inverse-consistent constraint to encourage that a pair of images are symmetrically deformed toward one another, and then propose an anti-folding constraint to minimize foldings in the estimated flow. The proposed method is evaluated on registration of T1-weighted brain MR images for tissue segmentation and anatomical landmark detection. Experimental results demonstrate that our ICNet outperforms several state-of-the-art algorithms for deformable image registration.

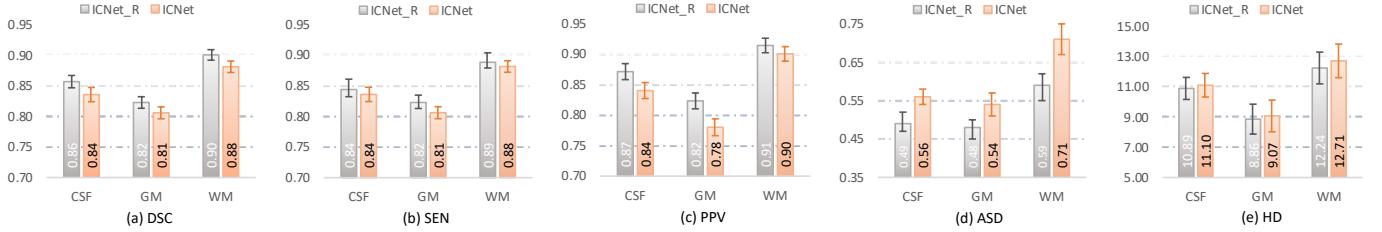


Fig. 10. Effect of using refine strategy for the proposed ICNet method. (a)-(e) show the tissue segmentation results in terms of five evaluation metrics. ICNet_R denotes ICNet with the network refining strategy.

In the current work, we utilize the mean squared distance (MSD) to measure the similarity between a warped source image and a target image, while many other similarity measures (such as correlation or mutual information) can be employed in the proposed deep-learning framework. Besides, only a simple network refining strategy is proposed in this work to handle the challenge of data heterogeneity (*e.g.*, in image contrast, resolution, and noise level), while more advanced data adaptation algorithms [71], [72] could be used to further boost the registration performance.

APPENDIX

As mentioned in the main text, besides using the conventional smoothness constraint [16], we also develop an anti-folding constraint to further prevent the folding in the learned discrete displacement field (*i.e.*, flow). Note that each element in a flow is a 3-dimensional vector (corresponding to three axes, *i.e.*, x , y , z), indicating the displacement of a particular voxel from its original location to a new location. We now explain the details of this anti-folding constraint defined in Eq. 5 in the main text.

We denote the to-be-estimated flow from an image **A** to an image **B** as \mathbf{F}_{AB} , and \mathbf{F}_{AB} is defined in the **A** space. For simplicity, we denote the 1-dimensional displacement in \mathbf{F}_{AB} along the i -th axis as \mathbf{F}_{AB}^i . As shown in Fig. 11, for the point m and its nearest neighbor point $m+1$ along the i -th axis in the space of **A**, we denote their displacements as $\mathbf{F}_{AB}^i(m)$ and $\mathbf{F}_{AB}^i(m+1)$, respectively.

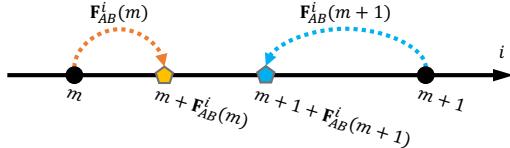


Fig. 11. Illustration of the proposed anti-folding constraint, with an index function to penalize the gradient along the i -th axis to avoid folding at the location m . Here, $m+1$ denotes the nearest neighbor of m . To avoid folding at the location of m , after the displacement $\mathbf{F}_{AB}^i(m)$ for m and $\mathbf{F}_{AB}^i(m+1)$ for $m+1$, the new locations of these two points should follow that $m + \mathbf{F}_{AB}^i(m) < m + 1 + \mathbf{F}_{AB}^i(m+1)$.

To avoid folding at the location of m , it is required that the new locations of these two points (after displacement via \mathbf{F}_{AB}^i) should follow

$$\begin{aligned} m + \mathbf{F}_{AB}^i(m) &< m + 1 + \mathbf{F}_{AB}^i(m+1) \\ \Rightarrow \mathbf{F}_{AB}^i(m+1) - \mathbf{F}_{AB}^i(m) &+ 1 > 0 \end{aligned} \quad (8)$$

where $m + \mathbf{F}_{AB}^i(m)$ is the new location of the point m , and $(m+1) + \mathbf{F}_{AB}^i(m+1)$ denotes the new location of the

point $m+1$. In discrete problems, the gradient of \mathbf{F}_{AB}^i at the location of m along the i -th axis is typically defined as

$$\begin{aligned} \nabla \mathbf{F}_{AB}^i(m) &= \frac{\mathbf{F}_{AB}^i(m+1) - \mathbf{F}_{AB}^i(m)}{(m+1) - m} \\ &= \mathbf{F}_{AB}^i(m+1) - \mathbf{F}_{AB}^i(m) \end{aligned} \quad (9)$$

Combining Eq. 9 to Eq. 8, we can get the following

$$\nabla \mathbf{F}_{AB}^i(m) + 1 > 0 \quad (10)$$

which indicates that, if $\nabla \mathbf{F}_{AB}^i(m) + 1 > 0$, there is no folding at the location of m . In contrast, if $\nabla \mathbf{F}_{AB}^i(m) + 1 \leq 0$, there is folding at the location of m .

Accordingly, to avoid folding in the learned flow, we propose an anti-folding constraint by penalizing the gradient of the flow, at locations that violate the rule in Eq. 8. Let $Q = \nabla \mathbf{F}_{AB}^i(m) + 1$, and we denote an index function $\delta(Q)$, where $\delta(Q) = |Q|$ if $Q \leq 0$, and $\delta(Q) = 0$, otherwise. Specifically, in the proposed anti-fold constraint (see Eq. 5 in the main text), if there is folding in the location of m along the i -th axis (*i.e.*, $\nabla \mathbf{F}_{AB}^i(m) + 1 \leq 0$), we enforce the penalty $|\nabla \mathbf{F}_{AB}^i(m) + 1|$ on the gradient at this location, requiring this gradient to be small. In contrast, if there is no folding at the location of m along the i -th axis (*i.e.*, $\nabla \mathbf{F}_{AB}^i(m) + 1 > 0$), we do not penalize the gradient of the estimated flow at this location.

REFERENCES

- [1] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.
- [2] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Landmark-based deep multi-instance learning for brain disease diagnosis," *Medical Image Analysis*, vol. 43, pp. 157–168, 2018.
- [3] C. Lian, J. Zhang, M. Liu, X. Zong, S.-C. Hung, W. Lin, and D. Shen, "Multi-channel multi-scale fully convolutional network for 3D perivascular spaces segmentation in TT MR images," *Medical Image Analysis*, vol. 46, pp. 106–117, 2018.
- [4] J. V. Hajnal and D. L. Hill, *Medical Image Registration*. CRC Press, 2001.
- [5] 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.
- [6] 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.
- [7] R. Bajcsy and S. Kovačić, "Multiresolution elastic matching," *Computer Vision, Graphics, and Image Processing*, vol. 46, no. 1, pp. 1–21, 1989.
- [8] 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.
- [9] 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.

- [10] X. Yang, R. Kwitt, and M. Niethammer, "Fast predictive image registration," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 48–57.
- [11] 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.
- [12] X. Cao, J. Yang, J. Zhang, Q. Wang, P.-T. Yap, and D. Shen, "Deformable image registration using cue-aware deep regression network," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1900–1911, 2018.
- [13] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of MR brain images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 649–656.
- [14] S. Miao, Z. J. Wang, and R. Liao, "A CNN regression approach for real-time 2D/3D registration," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1352–1363, 2016.
- [15] S. Shan, X. Guo, W. Yan, E. I. Chang, Y. Fan, Y. Xu *et al.*, "Unsupervised end-to-end learning for deformable medical image registration," *arXiv preprint arXiv:1711.08608*, 2017.
- [16] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Isgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 204–212.
- [17] 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.
- [18] S. Marius, K. Stefan, and P. Josien, P. W., "A rigidity penalty term for nonrigid registration," *Medical Physics*, vol. 34, no. 11, pp. 4098–4108, 1998.
- [19] H. Lester and S. R. Arridge, "A survey of hierarchical non-linear medical image registration," *Pattern Recognition*, vol. 32, no. 1, pp. 129–149, 1999.
- [20] L. G. H. Derek, G. B. Philipp, H. Mark, and J. H. David, "Medical image registration," *Physics in Medicine and Biology*, vol. 46, no. 3, pp. R1–R45, 2001.
- [21] Z. Barbara and F. Jan, "Image registration methods: A survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003.
- [22] M. Holden, "A review of geometric transformations for nonrigid body registration," *IEEE Transactions on Medical Imaging*, vol. 27, no. 1, pp. 111–128, 2008.
- [23] D. Rueckert and J. A. Schnabel, *Medical Image Registration*. Springer Berlin Heidelberg, 2011, pp. 131–154.
- [24] R. P. Woods, S. T. Grafton, C. J. Holmes, S. R. Cherry, and J. C. Mazziotta, "Automated image registration: I. General methods and intrasubject, intramodality validation," *Journal of Computer Assisted Tomography*, vol. 22, no. 1, pp. 139–152, 1998.
- [25] P. Hellier, J. Ashburner, I. Corouge, C. Barillot, and K. J. Friston, "Inter-subject registration of functional and anatomical data using SPM," in *Medical Image Computing and Computer-Assisted Intervention*. Springer Berlin Heidelberg, 2002, pp. 590–597.
- [26] 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.
- [27] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2008.
- [28] A. J., S. Suzi, and J. Mark, "FNIRT-FMRIB's non-linear image registration tool," *Human Brain Mapping*, 2008.
- [29] B. A. Ardekani, S. Guckemus, A. Bachman, M. J. Hoptman, M. Wojcieszek, and J. Nierenberg, "Quantitative comparison of algorithms for inter-subject registration of 3D volumetric brain MRI scans," *Journal of Neuroscience Methods*, vol. 142, no. 1, pp. 67–76, 2005.
- [30] D. L. Collins and A. C. Evans, "Animal: Validation and applications of nonlinear registration-based segmentation," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 11, no. 8, pp. 1217–1294, 1997.
- [31] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.
- [32] 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.
- [33] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35–51, 1996.
- [34] C. G. E. and J. H. J. F., "Consistent image registration," *IEEE Transactions on Medical Imaging*, vol. 20, no. 7, pp. 568–582, 2001.
- [35] R. D., A. P., H. R. A., H. J. V., and A. Hammers, "Diffeomorphic registration using B-splines," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2006, pp. 702–709.
- [36] O. Musse, F. Heitz, and J. P. Armsbach, "Topology preserving deformable image matching using constrained hierarchical parametric models," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1081–1093, 2001.
- [37] M. Sdika, "A fast nonrigid image registration with constraints on the Jacobian using large scale constrained optimization," *IEEE Transactions on Medical Imaging*, vol. 27, no. 2, pp. 271–281, 2008.
- [38] T. Mansi, X. Pennec, M. Serresant, H. Delingette, and N. Ayache, "iLogDemons: A demons-based registration algorithm for tracking incompressible elastic biological tissues," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 92–111, 2011.
- [39] C. Tanner, J. A. Schnabel, D. Chung, M. J. Clarkson, D. Rueckert, D. L. G. Hill, and D. J. Hawkes, "Volume and shape preservation of enhancing lesions when applying non-rigid registration to a time series of contrast enhancing MR breast images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2000.
- [40] W. Greene, S. Chelikani, K. Purushothaman, J. Knisely, Z. Chen, X. Papademetris, L. Staib, and J. Duncan, "Constrained non-rigid registration for use in image-guided adaptive radiotherapy," *Medical Image Analysis*, vol. 13, no. 5, pp. 809–817, 2009.
- [41] T. Rohlfing, C. R. Maurer, D. A. Bluemke, and M. A. Jacobs, "Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint," *IEEE Transactions on Medical Imaging*, vol. 22, no. 6, pp. 730–741, 2003.
- [42] A. Bustoquet, J. Oshinski, and O. Skrinjar, "Myocardial deformation recovery from cine MRI using a nearly incompressible biventricular model," *Medical Image Analysis*, vol. 12, no. 1, pp. 69–85, 2008.
- [43] J. Dauguet, A. Herard, J. Declerck, and T. Delzescaux, "Locally constrained cubic B-spline deformations to control volume variations," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 983–986.
- [44] D. Loeckx, F. Maes, D. Vandermeulen, and P. Suetens, "Nonrigid image registration using free-form deformations with a local rigidity constraint," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Springer Berlin Heidelberg, 2004, pp. 639–646.
- [45] J. Modersitzki, "FLIRT with rigidity-image registration with a local non-rigidity penalty," *International Journal of Computer Vision*, vol. 76, no. 2, pp. 153–163, 2008.
- [46] M. Lorenzi, N. Ayache, G. B. Frisoni, and X. Pennec, "LCC-Demons: A robust and accurate symmetric diffeomorphic registration algorithm," *NeuroImage*, vol. 81, pp. 470–483, 2013.
- [47] Y. Ou and C. Davatzikos, "DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting," *Medical Image Analysis*, vol. 15, no. 4, 2009.
- [48] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios, "Dense image registration through MRFs and efficient linear programming," *Medical Image Analysis*, vol. 12, no. 6, pp. 731–741, 2008.
- [49] Z. Xue, D. Shen, B. Karacali, J. Stern, D. Rottenberg, and C. Davatzikos, "Simulating deformations of MR brain images for validation of atlas-based segmentation and registration algorithms," *NeuroImage*, vol. 33, no. 3, pp. 855–866, 2006.
- [50] L. Wei, X. Cao, Z. Wang, Y. Gao, S. Hu, L. Wang, G. Wu, and D. Shen, "Learning-based deformable registration for infant MRI by integrating random forest with auto-context model," *Medical Physics*, vol. 44, no. 12, pp. 6289–6303, 2017.
- [51] M. Kim, G. Wu, P. Yap, and D. Shen, "A general fast registration framework by learning deformation-appearance correlation," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1823–1833, 2012.
- [52] Q. Wang, M. Kim, Y. Shi, G. Wu, and D. Shen, "Predict brain MR image registration via sparse learning of appearance and

- transformation," *Medical Image Analysis*, vol. 20, no. 1, pp. 61–75, 2015.
- [53] B. Gutiérrez-Becker, D. Mateus, L. Peter, and N. Navab, "Learning optimization updates for multimodal registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer International Publishing, 2016, pp. 19–27.
- [54] 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.
- [55] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration-A deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.
- [56] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "SVF-Net: Learning deformable image registration using shape matching," in *Medical Image Computing and Computer-Assisted Intervention*. Springer International Publishing, 2017, pp. 266–274.
- [57] H. Sokooti, B. de Vos, F. Berendsen, B. P. F. Lelieveldt, I. Işgum, and M. Staring, "Nonrigid image registration using multi-scale 3D convolutional neural networks," in *Medical Image Computing and Computer-Assisted Intervention*. Springer International Publishing, 2017, pp. 232–239.
- [58] 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 *Medical Image Computing and Computer-Assisted Intervention*.
- [59] A. Leow, S.-C. Huang, A. Geng, J. Becker, S. Davis, A. Toga, and P. Thompson, "Inverse consistent mapping in 3D deformable image registration: Its construction and statistical properties," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2005, pp. 493–503.
- [60] J. He and G. E. Christensen, "Large deformation inverse consistent elastic image registration," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2003, pp. 438–449.
- [61] 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.
- [62] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.
- [63] N. Ketkar, "Introduction to pytorch," in *Deep Learning with Python*. Springer, 2017, pp. 195–208.
- [64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [65] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowsky, P. J. Britson, J. L. Whitwell, and C. Ward, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.
- [66] C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans, "Enhancement of MR images using registration for signal averaging," *Journal of Computer Assisted Tomography*, vol. 22, no. 2, pp. 324–333, 1998.
- [67] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 45–57, 2001.
- [68] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, 2009.
- [69] J. Zhang, M. Liu, and D. Shen, "Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4753–4764, 2017.
- [70] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, and D. Shen, "Deformable image registration based on similarity-steered CNN regression," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 300–308.
- [71] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2288–2302, 2014.
- [72] I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2168–2175.