RSegNet: A Joint Learning Framework for Deformable Registration and Segmentation

Liang Qiu, *Student Member, IEEE*, and Hongliang Ren, *Senior Member, IEEE*

***Abstract*— Medical image segmentation and registration are two tasks to analyze the anatomical structures in clinical research. Still, deep-learning solutions utilizing the connections between segmentation and registration remain underdiscovered. This article designs a joint learning framework named RSegNet that can realize concurrent deformable registration and segmentation by minimizing an integrated loss function, including three parts: diffeomorphic registration loss, segmentation similarity loss, and dual-consistency supervision loss. The probabilistic diffeomorphic registration branch could benefit from the auxiliary segmentations available from the segmentation branch to achieve anatomical consistency and better deformation regularity by dual-consistency supervision. Simultaneously, the segmentation performance could also be improved by data augmentation based on the registration with well-behaved diffeomorphic guarantees. Experiments on the human brain 3-D magnetic resonance images have been implemented to demonstrate the effectiveness of our approach. We trained and validated RSegNet with 1000 images and tested its performances on four public datasets, which shows that our method successfully yields concurrent improvements of both segmentation and registration compared with separately trained networks. Specifically, our method can increase the accuracy of segmentation and registration by 7.0% and 1.4%, respectively, in terms of Dice scores.**

***Note to Practitioners*—Registration and segmentation of medical images are two significant tasks in medical research and clinical application. However, most existing approaches consider these two tasks independently while neglecting the potential association between them. Therefore, we suggest a new approach that combines these two tasks into one joint deep learning framework, boosting registration, and segmentation performance by introducing dual-consistency supervision. Besides, our framework could generate outputs within 1 s by taking an affinely aligned medical image pair as input, which is suitable for time-critical requirements in a clinic. We tested it on four public datasets and achieved state-of-the-art performance to demonstrate the proposed method's feasibility and robustness. Furthermore, our proposed RSegNet is a general learning framework suitable for various image modalities and anatomical structures. Hence, we expect our framework to serve as a practical clinical tool to speed up medical image analysis procedures and improve diagnostic accuracy.**

***Index Terms*— Deep learning, deformable registration, joint learning framework, segmentation.**

I. Introduction

ARTIFICIAL intelligence (AI) has been successfully applied in many healthcare areas, such as robotic surgery, virtual nursing assistance, clinical judgment or diagnosis, and medical image analysis, which has vast advances and perspectives in medical automation [1]–[3]. Typically, medical image registration and segmentation are two fundamental tasks in various medical image analyses, leading to many AI-powered tools and solutions. Segmentation aims at detecting the boundaries within a 2-D or 3-D image automatically or semiautomatically, which is often necessary to perform tasks, such as visual augmentation, computer-assisted diagnosis, interventions, and extraction of quantitative indices from images. A significant challenge of medical image segmentation is the high variability of medical images in various modalities and complicated anatomical characteristics. Traditional approaches usually perform segmentation employing edge detection filters or other mathematical methods with complex computation and lots of time consumption. Recently, deep learning techniques have been widely used in medical image segmentation and presented state-of-the-art performances but face many challenges [4]. For example, annotations for model training are usually expensive to obtain because expertise, effort, and time are required to produce precise class labeling, especially for the volumetric image comprised of a few 2-D slices.

Likewise, deformable registration is the process of establishing anatomical correspondence across different medical images acquired from different viewpoints, different modalities, and even at different times, which is a fundamental procedure for various clinical tasks, such as multimodality image fusion, tumor growth process monitoring, atlas-based segmentation, and organ template creation [5]. Deformable registration approaches usually include two stages: an initial low-dimensional rigid or affine transformation, followed by a much more complex deformable transformation that can capture the subtle and localized deformation. Even though extensive registration methods have been proposed, deformable

Manuscript received 29 March 2021; accepted 1 June 2021. Date of publication 22 June 2021; date of current version 5 July 2022. This article was recommended for publication by Associate Editor C. Park and Editor X. Xie upon evaluation of the reviewers' comments. This work was supported in part by the Shun Hing Institute of Advanced Engineering (SHIAE), The Chinese University of Hong Kong (CUHK), under Project BME-p1-21, 4720276. The work of Hongliang Ren was supported by the Singapore Academic Research Fund under Grant R397000353114. *(Corresponding author: Hongliang Ren.)*

Liang Qiu was with the Department of Biomedical Engineering, National University of Singapore, Singapore 119077. He is now with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: qiuliang@u.nus.edu).

Hongliang Ren is with the Department of Electronic Engineering, Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong, and also with the Department of Biomedical Engineering, National University of Singapore, Singapore 119077 (e-mail: hlren@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2021.3087868.

Digital Object Identifier 10.1109/TASE.2021.3087868

registration is still an active research area, especially when dealing with challenging medical image analysis tasks, e.g., large amounts of medical data analyses or comparison of multimodality images with significantly heterogeneous pathology. In view of the shortcomings of traditional deformable registration methods, such as intensive computation and local minimum problems, deep learning techniques can significantly improve the registration speed while guaranteeing competitive accuracy. Moreover, higher robustness can be acquired by deep learning-based algorithms due to its advantages of less concern about local optima, which can be bypassed because zero gradients often exist at saddle points [6].

However, most existing methods only conduct registration and segmentation separately and neglect the potential connection between these two tasks. Both registration and segmentation are correlated and can help to boost each other's performances. In this study, a joint learning framework called RSegNet is proposed for concurrent deformable registration and segmentation. Our contributions are summarized as follows:

1) We propose a joint learning framework by incorporating a probabilistic diffeomorphic registration network and a 3-D segmentation network. The probabilistic diffeomorphic registration branch could benefit from the auxiliary segmentation information available from the segmentation branch to achieve anatomical consistency and better deformation regularity by introducing consistency supervision. Simultaneously, the segmentation performance could also be enhanced by data augmentation based on the registration with well-behaved diffeomorphic guarantees.
2) Our proposed end-to-end framework can realize concurrent registration and segmentation around 1 s, suitable for time-critical tasks.
3) We evaluate RSegNet on four public datasets to show its generalizability and achieve consistent improvements on each dataset for both segmentation and registration tasks in terms of Dice scores compared with the separately learned networks.

The rest of the article is organized as follows. Section II introduces the related work. Section III provides background knowledge. Section IV presents the methods used in our proposed framework. Section V shows our experiments and analysis. Section VI describes the discussion and conclusion.

II. Related Work

A. Medical Image Segmentation

Medical image segmentation plays an increasingly significant role in current medical diagnosis, a procedure to detect the boundaries of organs or lesions within 2-D or 3-D images. However, the high variability of human anatomy and various medical imaging modalities have posed significant challenges to guarantee segmentation performances. Traditional segmentation approaches usually rely on the medical images' quantitative features or descriptors, such as intensity, shape, and texture, to distinguish the target anatomical structures. Besides, prior information on these image features is often involved in guiding the process of segmentation. Given the significant anatomical variability in medical images, extensive statistical inference models have been proposed to guarantee segmentation accuracy, such as the well-known active shape models [7] and active appearance models [8].

Moreover, atlas-based segmentation is also treated as a widely used paradigm that exploits the prior spatial knowledge in the atlas or template [9]–[15]. The atlas label information can be propagated to the subject image space with pixel/voxel correspondence by registration between the atlas and the subject image. Furthermore, deformable models have also been studied intensively, regarded as dominant medical image segmentation techniques [16], [17]. The deformable models are curves or surfaces that can conform to the target boundary guided by so-called internal and external forces. Although these classical segmentation algorithms can achieve relatively accurate results, they still suffer from complex computation and extensive time consumption.

Recently, deep learning for medical image segmentation and detection has comparable or superior performances versus experienced doctors [18], [19]. The existing deep learning methods for segmentation by exploiting the convolutional neural network (CNN) can be categorized into two groups: 2-D and 3-D CNN-based methods.

Instead of one-time processing of a whole volumetric image, 2-D CNN-based methods do it in a slice-by-slice way considering the computing capacity of commonly used graphic processing devices. One representative network is U-Net [20] that has been successfully applied in various medical image segmentation applications. U-Net's essential characteristic is the skip connection, which can provide deconvolution layers with significant high-resolution features among different stages of the network. Based on this structure, many variations have been proposed to deal with different segmentation tasks. For instance, a modified U-Net proposed in [21] was used to implement lung segmentation in X-ray scans, which showed a fast and accurate performance. Besides, an attention U-Net was proposed by integrating a new attention gate for medical imaging, which could learn how to focus on target anatomical structures [22]. Moreover, to further enhance the connection and reduce the semantic gap between encoder and decoder subnetworks, a nested U-Net called U-Net++ has been proposed, where the skip connection in each level has been redesigned using a series of nested dense convolutional blocks. Experiments have been carried out on different datasets, such as liver and chest CT scans, demonstrating U-Net++ with deep supervision achieved much better performances against U-Net [23].

Though 2-D CNN-based networks have improved the medical image segmentation performance compared with traditional methods, the spatial information embedded in the volumetric images is not fully exploited, affecting the segmentation accuracy. Then, 2.5-D CNN-based methods were proposed to exploit spatial information of neighboring pixels, providing slightly better performances over 2-D methods [24]. However, they are still limited by 2-D filters, which cannot extract powerful 3-D feature representation. Given all the limitations, several 3-D CNN-based approaches have been

provided. One representative is called 3-D U-Net [25], which extends the original U-Net into a 3-D version by successfully converting all 2-D operations into 3-D counterparts and densely segment volumetric images. Besides, a similar network V-Net [26] was proposed by incorporating residual connections, which can learn residuals within several stages to increase the accuracy and reduce the convergence time. By incorporating the novel Dice loss instead of traditional cross entropy, the imbalance problem existing between the foreground and the background during the training stage has been solved to some extent. However, all the methods above rely on manually annotated images and are susceptible to intensity variance. To tackle medical image segmentation with only a few labeled images, a semi-supervised, learning-based multiatlas segmentation strategy was proposed, which is further improved with spatial data augmentation and has been demonstrated to achieve higher Dice score and lower surface distance against supervised methods on anatomical structures of brain magnetic resonance (MR) images. Furthermore, a Bayesian brain MRI segmentation method was presented by exploiting a probabilistic atlas with deformable registration and obtaining outstanding segmentation accuracy on brain MR images [27].

B. Medical Image Registration

Classical medical image registration algorithms can provide impressive performances with rigorous and complex mathematical derivation to solve an optimization problem that can capture subtle, localized image deformation. Typically, almost all the traditional models exploit iterative optimization to control the local voxel correspondence given specific intensity-based similarity metrics, such as mean square distance (MSD) [28], (normalized) cross correlation (CC) [29], and (normalized) mutual information (MI) [30]–[32]. To guarantee the topology-preserving attribute, a large number of diffeomorphic registration algorithms have been proposed, such as widely used diffeomorphic demons [33], SyN [34], and LDDMM [35]. However, these traditional methods have several drawbacks. First, the registration process is computationally costly and time-consuming for each pair of images because of the high-dimensional iterative optimization strategies of these methods although the runtime can be reduced to several minutes even with graphic processing units (GPUs) to accelerate the computation. Second, the optimization is probably stuck in a local minimum point due to the objective function's nonconvexity property. Third, the registration performance may deteriorate when the algorithms are implemented on subject images with considerable anatomical variations. Therefore, efficient and robust medical image registration approaches are essential in today's computer-assisted medicine.

Compared with traditional registration approaches, deep learning techniques are powerful tools that are well-suited for medical image registration regarding computational efficiency and accuracy. Different deep learning algorithms have been proposed for both unimodal and multimodal registrations, in which the medical image data are usually 3-D computed tomography (CT), MR imaging (MRI), ultrasound imaging (US), cone-beam CT (CBCT), and 2-D X-ray images. For example, a Quicksilver registration algorithm with an encoder–decoder network was proposed to predict the deformation field of 2-D/3-D brain MR images [36]. Furthermore, a multiscale 3-D CNN named RegNet was designed to predict the deformation of 3-D chest CT data with the artificially generated displacement vector field (DVF) [37]. Contextual information is considered to boost the feature representation by integrating multiscaled MRI contents. Though those above deep learning-based registration methods have presented competitive registration performances, they are all supervised. To be more specific, ground-truth warp fields should be provided to guarantee practical training, usually from two sources. One is known transformations to compute synthetic medical images, which usually contain artificial noise. It may make the trained network less robust when processing challenging real images. The other is the deformation fields generated by specific traditional registration algorithms, which may bring in biases.

Given the rarely available ground-truth warp fields and their potential problems, unsupervised learning methods have been adopted for medical image registration. For example, a fully convolutional network named DIRNet [38] was provided to realize the unsupervised learning of the deformable image registration of cardiac cine MRI scans with normalized CC as the similarity metric, which has a comparable or slightly better performance than SimpleElastix [39], a conventional registration approach. Since DIRNet is only applicable to 2-D images, de Vos *et al.* [40] made an extension from 2-D to 3-D images and used multistage, multiresolution strategies to realize the unsupervised affine and deformable image registration with the B-spline transformation model. Recently, anatomical segmentations have been incorporated into the objective function to learn the 3-D voxel correspondence, indicating an accurate registration performance [41]. The authors employed task-specific higher-level label correspondences, including solid organs, ducts, vessels, and other *ad hoc* structures instead of traditional intensity-based similarity metrics to optimize the global and local displacement cross-modality registration network jointly.

Similarly inspired, Balakrishnan *et al.* [42] proposed a general unsupervised deformable registration network suitable for both unimodal and multimodal theoretically using MSD and CC as image similarity metrics with the help of segmentation-based loss, which achieves competitive accuracy compared with several classic well-performed methods, including SyN [43] and NiftiReg [44]. Furthermore, considering that the intensity-based or label-driven image similarity metrics may not perform very well in challenging registration tasks, several deep adversarial image registration approaches were proposed to achieve high-accurate registration, which uses discriminators to judge the divergence between moving and fixed images [45]–[48]. However, although using the generative adversarial network (GAN) [49] framework could realize deformable image registration, it may still suffer from common drawbacks, such as mode collapse and failure to converge. In general, although all the unsupervised methods mentioned above formulate the image registration as

optimization problems based on different similarity metrics and regularization terms, they cannot guarantee the diffeomorphism. To alleviate this problem, an unsupervised inference algorithm was derived based on a probabilistic generative model using diffeomorphic deformation representation [50]. In addition, anatomical surfaces are leveraged to improve the registration performance of the proposed framework, which can also guarantee its diffeomorphic property and fast runtime.

C. Joint Optimization of Segmentation and Registration

Although extensive approaches have been developed to solve segmentation and registration as independent problems, they are highly correlated. This section will further discuss the interdependence of segmentation and registration and introduce the motivation of the joint optimization of these two. Based on the literature review above, we can find that these two tasks can boost their performances by exploiting the auxiliary anatomical information or transformation from each other. For example, learning-based registration can be improved by incorporating the auxiliary segmentation information into the training procedure, which can be called weakly-supervised registration and demonstrated to help register [41], [42], [51]. Likewise, segmentation can be accomplished by registering the label information from atlas space to subject image space by geometric transformation called label propagation, widely known as atlas-based segmentation, such as those classical iterative optimization algorithms [9]–[15]. However, the ultimate segmentation results rely on the atlas-to-target registration performance, while the registration procedure, which only depends on an intensity-based similarity measure, usually suffers from the lack of anatomical consistency, such as shape or texture structure.

To overcome these limitations, alternative or joint optimization of both segmentation and registration has been investigated. Specifically, the alternative registration and segmentation method means one-step registration followed by one-step segmentation, which assists the counterpart in an iterative interleaving way [52], [53]. For example, a generative probabilistic framework was proposed in [52], which combines registration, classification, and bias correction in a circular procedure by establishing a maximum a posteriori model. The model parameters are estimated alternatively for each component using an expectation–maximization (EM) algorithm. Likewise, joint registration and segmentation approaches update the parameters simultaneously in each iteration, considering more local constraints to obtain a deformation field. A joint registration and segmentation model derived from optical flow and active contour theory was proposed for automatic subthalamic nucleus targeting on MRI [54]. It successfully combines the nonrigid registration model's strength with an active contour framework embedded with local segmentation constraints, which can theoretically register any type of anatomical contours. Some other methods try to calculate the deformation field between corresponding anatomical contours extracted from atlas and subject images, respectively, which can be further classified into two categories: the energy-based variational framework [55] and the Markov random field-based methods [56], [57].

However, those traditional joint optimization methods need complex mathematical modeling and intensive computation for each image pairs, which are not suitable for some time-critical clinical diagnosis or preoperative analysis. To cope with this problem, deep learning-based joint optimization methods have been exploited. An unsupervised atlas-based segmentation method is proposed recently, which jointly learns the registration and segmentation and achieves a remarkable brain MRI segmentation performance [27]. Nevertheless, the deformation field is learned from the registration performed on a prior probabilistic atlas map. It is a single-task model only for segmentation while not exactly for multiple tasks simultaneously. Moreover, a joint semi-supervised learning framework named DeepAtlas was proposed by optimizing the segmentation and registration branch network alternatively [58], which provides a general solution to cope with joint training when many images are with few manual segmentations. The imperfect supervision can assist one subnetwork with unlabeled data from the other subnetwork by introducing an anatomical similarity loss. Experiments performed on the knee and brain MRIs have demonstrated its effectiveness compared to separately learned networks, with manual segmentations available for all the training images to provide the upper bound. Furthermore, an unsupervised probabilistic model named U-RSNet was proposed to realize concurrent medical image registration and segmentation in one framework, without the need for any known transformations and manually segmented images, which can efficiently improve the data processing procedure and has been demonstrated effective on various brain MRI datasets [59].

This article proposes a novel joint learning framework for concurrent registration and segmentation with dual-consistency supervision integrated by the invertible deformation field with a well-behaved diffeomorphic guarantee, whose performance could be improved compared with separately trained networks, when all the segmentation ground-truth data are available.

III. Weakly-Supervised Deformable Registration

A. Image-Based Deformable Registration

Deformable registration is a nonrigid registration procedure to establish the voxelwise correspondence between the moving and fixed images with an estimated deformation field. Let F and M be the fixed image and the moving image, respectively, and let $\boldsymbol{\phi}$ denote the deformation field; then, the typical optimization problem for deformable registration can be formulated as

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} \left[\mathcal{L}_{\text{sim}}(F, M \circ \boldsymbol{\phi}) + \lambda_{\text{reg}} \mathcal{L}_{\text{smooth}}(\boldsymbol{\phi})\right] \tag{1}$$

where $\boldsymbol{\phi}^*$ represents the optimal deformation field, $\mathcal{L}_{\text{sim}}(\cdot, \cdot)$ denotes the similarity loss to penalize the dissimilarity between the image F and the warped image $M \circ \boldsymbol{\phi}$, $\mathcal{L}_{\text{smooth}}(\cdot)$ is the regularization loss to encourage the smoothness deformation field $\boldsymbol{\phi}$, and λ_{reg} is the tradeoff hyperparameter.

Many deformable registration approaches parameterize the deformable transformation $\boldsymbol{\phi}$ using a displacement field u,

which is simply added to an identity transformation x, shown as follows:

$$\boldsymbol{\phi}(x) = x + u(x). \tag{2}$$

With such intuitive parameterization, the inverse transformation can be approximated by subtracting the displacement. However, the precise inverse transformation may not exist, especially for large deformation [60]. To guarantee the topology-preserving of the deformation field $\boldsymbol{\phi}$, diffeomorphic registration is exploited to provide differentiable and invertible deformations, which can be formulated as an ordinary differential equation as follows:

$$\frac{\partial \boldsymbol{\phi}^{(t)}}{\partial t} = \boldsymbol{v}(\boldsymbol{\phi}^{(t)}) \tag{3}$$

where $\boldsymbol{\phi}^{(0)}$ indicates the identity transformation and the deformation field $\boldsymbol{\phi}^{(1)}$ can be acquired by the integration of the corresponding velocity filed $\boldsymbol{v}$ over $t = [0, 1]$. According to group theory, $\boldsymbol{\phi}^{(1)} = \exp(\boldsymbol{v})$ can be obtained by recursively computing $\boldsymbol{\phi}^{(1/2^{t-1})} = \boldsymbol{\phi}^{(1/2^t)} \circ \boldsymbol{\phi}^{(1/2^t)}$ T times with an initial condition $\boldsymbol{\phi}^{(1/2^T)} = p + \boldsymbol{v}/2^T$ (p is a map of spatial locations) implemented by the scaling and squaring network [61].

B. Weakly-Supervised Deformable Registration

Apart from directly modeling the deformable registration as an optimization problem with only intensity images, such as formula (1), anatomical segmentation maps can also be leveraged as auxiliary information to improve the registration performance, called weakly-supervised or semi-supervised deformable registration. It is first proposed by Hu *et al.* [41] and then further expanded and discussed in [42], whose core idea is that, if the estimated deformation field $\boldsymbol{\phi}$ is accurate enough, the specific regions in F and $M \circ \boldsymbol{\phi}$ corresponding to the same anatomical structure should overlap as well. Therefore, an additional segmentation loss $\mathcal{L}_{\text{seg}}(\cdot, \cdot)$ to encourage the anatomical structure consistency is added to formula (1), shown as follows:

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} \begin{bmatrix} \mathcal{L}_{\text{sim}}(F, M \circ \boldsymbol{\phi}) + \lambda_{\text{reg}} \mathcal{L}_{\text{smooth}}(\boldsymbol{\phi}) \\ +\lambda_{\text{seg}} \mathcal{L}_{\text{seg}}(S_F, S_M \circ \boldsymbol{\phi}) \end{bmatrix} \tag{4}$$

where λ_{seg} is the tradeoff hyperparameter, and S_F and S_M represent the corresponding anatomical structures in image F and M. Because $\mathcal{L}_{\text{seg}}(\cdot, \cdot)$ cannot encourage the smoothness of the deformation field, λ_{reg} and λ_{seg} should be carefully determined to guarantee a reasonable registration performance.

IV. Method

A. System Overview

Given the potential connection of image registration and segmentation, our goal is to develop a joint learning framework for both tasks simultaneously, improving registration and segmentation accuracy. Unlike the weakly-supervised deformable registration mentioned above, which directly exploits auxiliary annotated segmentation to improve single-task performance, our framework comprises two independent subnetworks, an unsupervised registration network and a supervised segmentation network, connected by consistency supervision to achieve bidirectional improvements.

Let m and f represent the moving 3-D image and the fixed 3-D image, respectively. All the 3-D images are assumed to be affinely aligned, with only nonlinear misalignment left.

The overview of our approach is presented in Fig. 1. Our proposed neural network's whole architecture comprises two branches, which correspond to the registration subnetwork Reg-SubNet and the segmentation subnetwork Seg-SubNet, respectively. For image registration, we mainly adopt the probabilistic model to guarantee the diffeomorphic characteristic. The registration 3-D U-Net $\mathcal{R}_\theta(m, f)$ from the registration subnetwork Reg-SubNet takes a moving 3-D image m and a fixed 3-D image f as inputs and outputs a probabilistic velocity field $\boldsymbol{v}$ sampled from corresponding mean $\boldsymbol{\mu}_{\boldsymbol{v}|m,f}$ and variance $\sum_{\boldsymbol{v}|m,f}$. Using squaring and scaling integration layers, the velocity field $\boldsymbol{v}$ is transformed into a diffeomorphic deformation field $\boldsymbol{\phi}_{\boldsymbol{v}}$, which can transform m to obtain $m \circ \boldsymbol{\phi}_{\boldsymbol{v}}$ to optimally fit f by the spatial transformation network (STN) [62], i.e., for $\forall x \in R^3$, $m \circ \boldsymbol{\phi}_{\boldsymbol{v}}(x)$ and $f(x)$ correspond to the similar anatomical location, namely, $[m \circ \boldsymbol{\phi}_{\boldsymbol{v}}](x) \approx f(x)$. The deformation field $\boldsymbol{\phi}_{\boldsymbol{v}}$ between the moving image m and the fixed image f can be treated as a 3-D image with three channels, representing corresponding displacement components, respectively. For image segmentation, the segmentation 3-D U-Net $\mathcal{S}_\vartheta(m)$ takes image m as input and outputs a segmented 3-D image $\mathcal{S}_m$. This branch is initially trained by comparing the segmented 3-D image $\mathcal{S}_m$ with the labeled moving 3-D segmentation S_m by minimizing a soft multiclass Dice loss to find optimal parameters of the segmentation network.

Given these two tasks' respective characteristics, they can be highly correlated and improve each other's performances. Segmentation can provide auxiliary anatomical structure information to help the registration algorithm find regional correspondence, and conversely, the segmentation task can also benefit from registration with data augmentation. Thus, in the network architecture, a dual-consistency module is designed to connect these two tasks, as illustrated in the light blue window of Fig. 1. To integrate the dual-consistency structure into the whole framework, a probabilistic diffeomorphic registration is exploited in our framework, providing the invertible deformation field with a well-behaved diffeomorphic guarantee. Minimizing both $\mathcal{L}_{\text{forward-cons}}$ and $\mathcal{L}_{\text{inverse-cons}}$ can provide consistent supervision to guide the training process. The detailed network architecture and loss functions will be described in the following.

B. Diffeomorphic Registration

In our framework, we adopt an unsupervised registration method as our Reg-SubNet. To further ensure certain desirable registration properties, we build Reg-SubNet based on a probabilistic diffeomorphic registration method VM-diff [50], which can provide a differentiable, invertible topology-preserving deformation field for image deformable transformation. The CNN $\mathcal{R}_\theta(m, f)$ parameterized by θ to estimate voxelwise mean $\boldsymbol{\mu}_{\boldsymbol{v}|m,f}$ and variance $\sum_{\boldsymbol{v}|m,f}$ can be trained by the

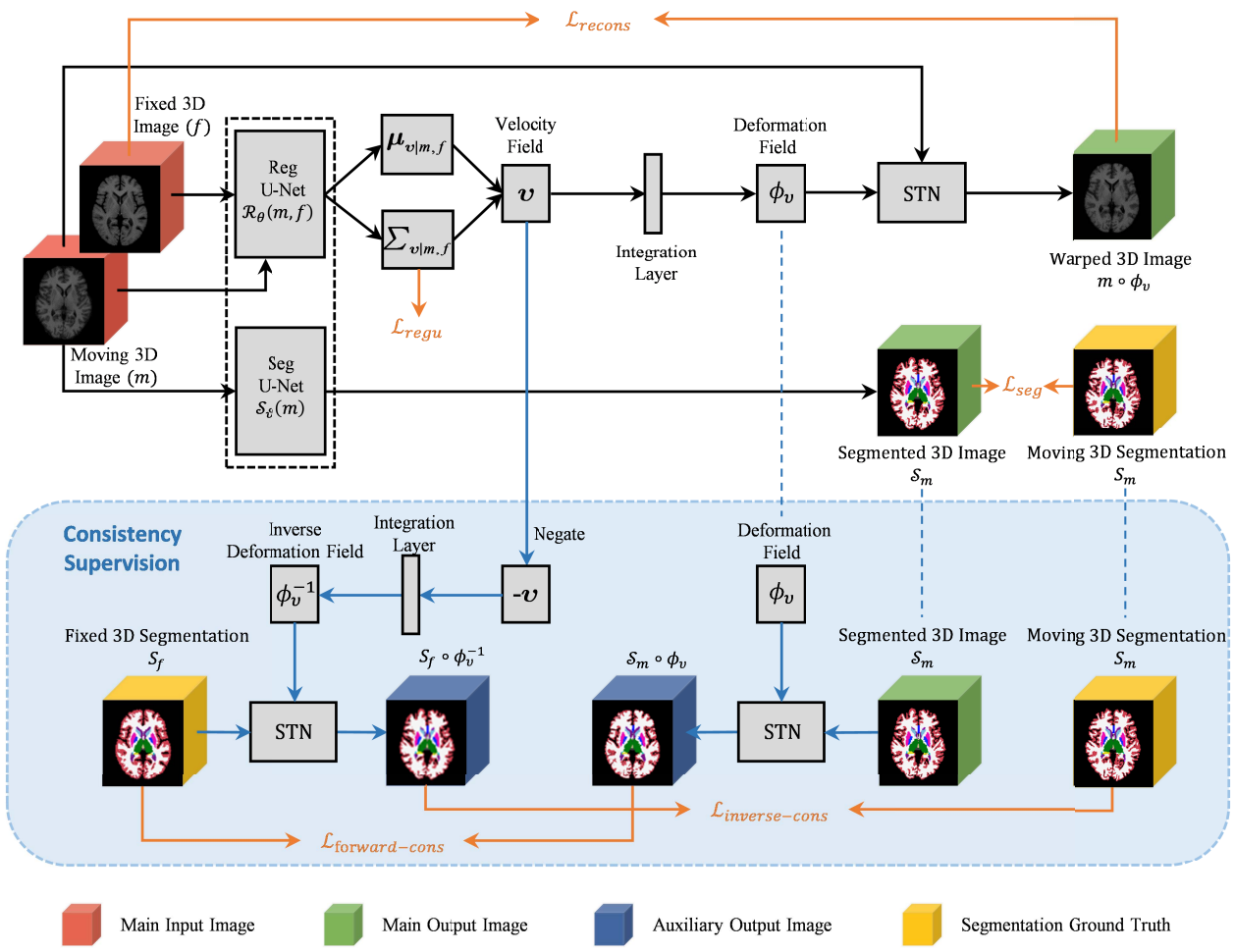

Fig. 1. Overview of RSegNet, which is a joint learning framework for both image registration and segmentation. Our framework is mainly composed of two branches: registration subnetwork and segmentation subnetwork. The 3-D enhanced U-Net $\mathcal{R}_\theta(m, f)$, a part of registration subnetwork Reg-SubNet, takes a moving 3-D image m and a fixed 3-D image f as inputs and outputs a probabilistic velocity field $\boldsymbol{v}$ sampled from corresponding mean and variance. By using squaring and scaling integration layers, the velocity field $\boldsymbol{v}$ is transformed into a diffeomorphic deformation field $\boldsymbol{\phi_v}$, which can warp m to obtain $m \circ \boldsymbol{\phi_v}$ by STN. Customized $\mathcal{S}_\vartheta(m)$ that is the segmentation subnetwork Seg-SubNet takes image m as input and outputs a segmented 3-D image $\mathcal{S}_m$. The light blue window illustrates the computation of the dual-consistency supervision, which can be perfectly integrated into the whole framework with the invertible deformation field produced by the introduced probabilistic diffeomorphic registration. Minimizing both $\mathcal{L}_{\text{forward-cons}}$ and $\mathcal{L}_{\text{inverse-cons}}$ can provide a consistency guarantee to guide the training process. For better illustration, the main input images are labeled in red, the main output images are labeled in green, and the segmentation ground-truth images for loss calculation are highlighted in yellow. Moreover, the auxiliary output images for consistency supervision, which are only used during the training stage, are highlighted in dark blue. Notably, the fixed 3-D segmentation ground truth is also exploited as an auxiliary input image during the joint training procedure.

approximate posterior inference

$$L(\theta|f,m) = -\mathrm{E}_q\big[\log p(f|\boldsymbol{v},m)\big] + \mathrm{KL}\big[q_\psi(\boldsymbol{v}|f,m)\|p(\boldsymbol{v})\big] \tag{5}$$

whose rigorous theoretical derivation can be easily found in [50]. After obtaining $\boldsymbol{\mu}_{\boldsymbol{v}|m,f}$ and $\sum_{\boldsymbol{v}|m,f}$, the stationary velocity field $\boldsymbol{v}$ could be generated by resampling from a normal distribution $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{v}|m,f}, \sum_{\boldsymbol{v}|m,f})$, which can be further processed into a deformation field $\boldsymbol{\phi_v}$.

The particular architecture of $\mathcal{R}_\theta(m, f)$ in our experiments is quite similar to 3-D U-Net [25], including encoder, decoder, and direct skip connection. Besides, related attributes have been adjusted to fit our application, such as the number of the convolution kernel and the architecture's depth, as shown in Fig. 2. It should be noted that $\mathcal{R}_\theta(m, f)$ is a general CNN block built with different architectures to realize similar feature extraction and provide ideal output, but the specific design is not the focus of this article.

In the encoder stage of $\mathcal{R}_\theta(m, f)$, 3-D convolutions are applied to obtain increasingly abstract features of the original input images as the process goes deeper. The kernel size and the stride in each convolution are set to 3 and 2, respectively, followed by a LeakyReLU. The vertical depth of the 3-D U-Net can be called level, the spatial dimension is reduced in half in each subsequent level, and the number of the feature channels is increased to 32 and 64. In the decoder stage, upsampling layers are used to recover the dimension to half of the input size to get $\boldsymbol{\mu}_{\boldsymbol{v}|m,f}$ and $\sum_{\boldsymbol{v}|m,f}$, and concatenation layers are utilized to recover the lost details due to dimensionality reduction in the encoder stage by concatenating the corresponding feature maps of the encoder and the decoder from the same level.

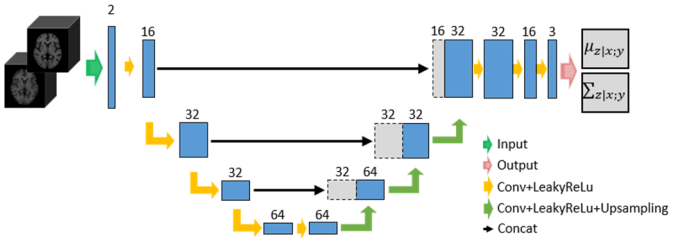

Fig. 2. Architecture of diffeomorphic registration 3-D U-Net takes the 3-D image m and f as input and outputs voxelwise mean $\boldsymbol{\mu}_{\boldsymbol{v}|m,f}$ and variance $\sum_{\boldsymbol{v}|m,f}$ of the stationary velocity field $\boldsymbol{v}$.

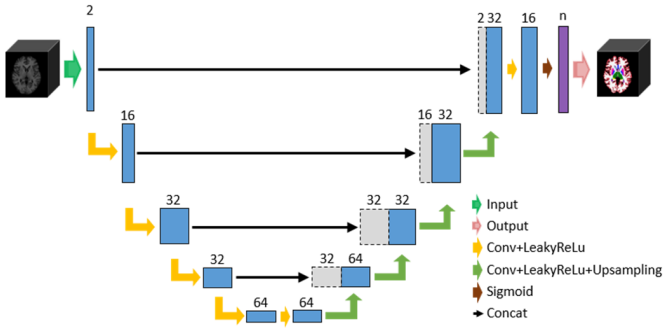

Fig. 3. 3-D U-Net architecture takes a 3-D image as input and outputs the corresponding segmentation with n different ROIs.

C. 3-D Image Segmentation

$\mathcal{S}_{\vartheta}(m)$ is the core architecture of the segmentation subnetwork Seg-SubNet, which takes a moving image m as input and outputs corresponding segmentation with n regions of interest (ROIs). It should be noted that various segmentation CNNs can achieve similar performance, and the exact design $\mathcal{S}_{\vartheta}(m)$ is not the focus in our RSegNet framework. In view of the excellent performance of 3-D U-Net [25] widely applied in medical image segmentation, we design $\mathcal{S}_{\vartheta}(m)$ with a similar architecture, as shown in Fig. 3.

In light of the similar characteristic attributes of the input and the output and the limited GPU computing resources, we exploit the similar structure of $\mathcal{R}_{\theta}(m, f)$ as the segmentation architecture in our experiments. Instead of outputting a half-dimensional stationary velocity field, the upsampling layer is applied at the end of the network to recover the resolution to the original size of the input images. A sigmoid activation layer is followed to output the predictive probabilistic distributions of n ROIs in each voxel. In the training stage, various supervised segmentation loss functions can optimize the segmentation network, such as multiclass Dice loss. More details about the loss function will be described in Section IV-E.

D. Consistency Supervision

In Sections IV-B and IV-C, registration and segmentation subnetworks in our RSegNet have been introduced, which can complete their own tasks, respectively. However, they are closely related and can help each other to improve their performance. The segmentation result $\mathcal{S}_m$ produced by $\mathcal{S}_{\vartheta}(m)$ provides the probabilistic distributions of n ROIs, assigning each voxel of the input image to an anatomical structure. If the deformation field $\boldsymbol{\phi}_{\boldsymbol{v}}$ generated by $\mathcal{R}_{\theta}(m, f)$ is accurate, then $\mathcal{S}_m \circ \boldsymbol{\phi}_{\boldsymbol{v}}$ and the fixed 3-D segmentation S_f should have an ideal overlap of corresponding anatomical structure distributions, which could be implemented by optimizing the loss $\mathcal{L}_{\text{forward-cons}}$. That is to say, this constraint can help the registration algorithm to find accurate regional correspondence, and it can also enhance the segmentation performance with an additional data augmentation.

Furthermore, in the network architecture, the negative velocity field $-\boldsymbol{v}$ can be easily integrated to compute the inverse deformation field $\boldsymbol{\phi}_{\boldsymbol{v}}^{-1}$ due to the diffeomorphic guarantee, which can be applied to the fixed 3-D segmentation S_f to obtain $S_f \circ \boldsymbol{\phi}_{\boldsymbol{v}}^{-1}$. Ideally, $S_f \circ \boldsymbol{\phi}_{\boldsymbol{v}}^{-1}$ and the moving 3-D segmentation S_m should be as similar as possible, making the registration subnetwork benefit from the auxiliary segmentation information directly to achieve anatomical consistency with loss $\mathcal{L}_{\text{inverse-cons}}$, which is a weakly-supervised registration used with the inverse form of the deformation field. This auxiliary term will enhance the registration task as well, whose improvement will further benefit the segmentation performance, inversely.

In general, we define the bidirectional constraint conditions $\mathcal{L}_{\text{forward-cons}}$ and $\mathcal{L}_{\text{inverse-cons}}$ as the dual-consistency supervision $\mathcal{L}_{\text{cons}}$, which establishes the connection between registration and segmentation and provides proper guidance for the parameter updating of both tasks during a single training procedure. Ascribed to the invertible deformation field generated by the probabilistic diffeomorphic registration [50], this consistency module can be perfectly integrated into one framework to form our complete RSegNet.

E. Loss Function and Implementation Details

This section proposes a loss function for the joint learning framework RSegNet, which consists of three components: $\mathcal{L}_{\text{reg}}$ that indicates the loss of registration subnetwork, $\mathcal{L}_{\text{seg}}$ that denotes the loss of segmentation subnetwork, and $\mathcal{L}_{\text{cons}}$ that represents the loss of consistency supervision.

The registration loss $\mathcal{L}_{\text{reg}}$ includes two parts according to the probabilistic model VM-diff [50] shown in formula (5): the reconstruction loss $\mathcal{L}_{\text{recon}}$ that encourages image $m \circ \boldsymbol{\phi}_{\boldsymbol{v}}$ and image f as similar as possible and the regulation loss $\mathcal{L}_{\text{regu}}$ that encourages the posterior $q_{\theta}(\boldsymbol{v}|m, f)$ to be close to the prior $p(\boldsymbol{v})$, whose entire derivation can be found in [50], shown as follows:

$$\mathcal{L}_{\text{recon}} = \frac{1}{2\sigma^2 K} \sum_{k} \left\| f - m \circ \boldsymbol{\phi}_{\boldsymbol{v}_k} \right\|^2 \tag{6}$$

$$\mathcal{L}_{\text{regu}} = \left[\boldsymbol{\mu}_{\boldsymbol{v}|m;f}^{T} \boldsymbol{\Gamma}_{\boldsymbol{v}} \boldsymbol{\mu}_{\boldsymbol{v}|m;f} + \text{tr}\left(\lambda \boldsymbol{D} \sum_{\boldsymbol{v}|m,f} - \log \sum_{\boldsymbol{v}|m,f} \right) \right] \Big/ 2 \tag{7}$$

where K is the sampling number, $\boldsymbol{\Gamma}$ and $\boldsymbol{D}$ are the precision matrix and the graph degree matrix, respectively, and λ is the parameter to control the scale of $\boldsymbol{v}$.

In addition, the segmentation loss $\mathcal{L}_{\text{seg}}$ used in our experiments is the multiclass Dice loss to guarantee the

differentiation, shown as follows:

$$\mathcal{L}_{\text{seg}} = -\frac{1}{N}\sum_{n=1}^{N}\frac{\sum_x \mathcal{S}_m(x)\cdot S_m(x)}{\sum_x \mathcal{S}_m(x)+\sum_x S_m(x)} \tag{8}$$

where N indicates the number of the ROIs to be segmented and x represents the voxel location. $\mathcal{S}_m(x)$ and $S_m(x)$ denote the prediction and the ground truth of segmentation at voxel x, respectively.

Similarly, the loss of our dual-consistency supervision $\mathcal{L}_{\text{cons}}$ contains two components, namely, $\mathcal{L}_{\text{forward-cons}}$ and $\mathcal{L}_{\text{inverse-cons}}$, shown as follows:

$$\mathcal{L}_{\text{forward-cons}} = -\frac{1}{N}\sum_{n=1}^{N}\frac{\sum_x [\mathcal{S}_m(x)\circ \boldsymbol{\phi}_{\boldsymbol{v}}]\cdot S_f(x)}{\sum_x [\mathcal{S}_m(x)\circ \boldsymbol{\phi}_{\boldsymbol{v}}]+\sum_x S_f(x)} \tag{9}$$

$$\mathcal{L}_{\text{inverse-cons}} = -\frac{1}{N}\sum_{n=1}^{N}\frac{\sum_x \left[S_f(x)\circ \boldsymbol{\phi}_{\boldsymbol{v}}^{-1}\right]\cdot S_m(x)}{\sum_x \left[S_f(x)\circ \boldsymbol{\phi}_{\boldsymbol{v}}^{-1}\right]+\sum_x S_m(x)} \tag{10}$$

where $\mathcal{S}_m(x)\circ \boldsymbol{\phi}_{\boldsymbol{v}}$ and $S_f(x)\circ \boldsymbol{\phi}_{\boldsymbol{v}}^{-1}$ indicate the warped segmentation by the mutually invertible deformation fields $\boldsymbol{\phi}_{\boldsymbol{v}}$ and $\boldsymbol{\phi}_{\boldsymbol{v}}^{-1}$, respectively. In addition, $S_f(x)$ and $S_m(x)$ indicate the segmentation of the fixed 3-D image and the moving 3-D image, which are treated as ground truths. Notably, to guarantee the differentiation of the optimization based on stochastic gradient descent, $S_m(x)$ and $S_f(x)$ are transformed to the volumes with N channels, each of which is a binary mask that corresponds to a particular anatomical label.

Then, we can combine all the loss items to obtain the objective loss function

$$\begin{aligned}\mathcal{L}_{\text{RSegNet}} &= \mathcal{L}_{\text{reg}} + \alpha\mathcal{L}_{\text{seg}} + \beta\mathcal{L}_{\text{cons}} \\ &= (\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{regu}}) + \alpha\mathcal{L}_{\text{seg}} \\ &\quad +(\beta_1\mathcal{L}_{\text{forward-cons}} + \beta_2\mathcal{L}_{\text{inverse-cons}})\end{aligned} \tag{11}$$

where α, β_1, and β_2 represent the hyperparameters.

Since the imbalanced learning speed of the registration and segmentation subnetwork is due to different task characteristics and model complexities, optimizing the whole framework straightforward may not work well at the beginning of the training stage. In terms of the specific architecture used in our experiments, the segmentation subnetwork converges much faster than the registration subnetwork. Instead of joint training with consistency supervision from scratch, we can pretrain the two subnetworks separately in a simple way by only setting the hyperparameters β_1 and β_2 to zero until both subnetworks achieve satisfactory and stable performances. After that, the whole joint learning framework could be trained by minimizing the objective loss function (11) with fine-tuned hyperparameters α, β_1, and β_2.

To predict accurately by optimizing the cost function shown in (11), the data going through the prediction stage must have a similar distribution as those on which the model was trained. Otherwise, the predictive performance may deteriorate due to model drift. Since training is much more computationally demanding than inferring, retraining based on the available model learned from the generalized or similar dataset will quickly refine the model weights to adapt to the target data domain and achieve good performance. It should be noted that retraining simply refers to rerunning the deep-learning model on a new training dataset without algorithm modification, which is also called transfer learning under some circumstances.

TABLE I
ROIS IN OUR DATASET

ID	ROI name	ID	ROI name
1	Left Cerebral White Matter	16	Right Cerebral White Matter
2	Left Cerebral Cortex	17	Right Cerebral Cortex
3	Left Lateral Ventricle	18	Right Lateral Ventricle
4	Left Cerebellum white Matter	19	Right Cerebellum white Matter
5	Left Cerebellum Cortex	20	Right Cerebellum Cortex
6	Left Thalamus Proper	21	Right Thalamus Proper
7	Left Caudate	22	Right Caudate
8	Left Putamen	23	Right Putamen
9	Left Pallidum	24	Right Pallidum
10	Left Hippocampus	25	Right Hippocampus
11	Left Amygdala	26	Right Amygdala
12	Left VentralDC	27	Right VentralDC
13	Left Choroid Plexus	28	Right Choroid Plexus
14	3rd Ventricle	29	4rd Ventricle
15	CSF	30	Brain Stem

V. EXPERIMENTS

A. Medical Image Data and Preprocessing

Many brain images are needed to train our joint learning-based RSegNet, just like other deep learning networks. We randomly collect 1000 brain MRI training images with various age ranges and health conditions from four public datasets, OASIS [63], ADNI [64], ABIDE [65], and ADHD200 [66], where 800 images for training and 200 images for validation, and test our approach on another four different datasets: CUMC12 [67], IBSR18 [67], MGH10 [67], and LPBA40 [68]. For all the raw data mentioned above, preprocessing should generate suitable images to feed into our RSegNet. Specifically, skull stripping will firstly be applied to all the datasets to get corresponding linearly resampled brain masks ($256 \times 256 \times 256$ resolution) with 1 mm isotropic voxels using FreeSurfer [69], followed by image normalization, image cropping, and manually refinement, which results in well-defined images with $160 \times 192 \times 224$ resolution. Moreover, all the brain images are segmented into 30 specific anatomical structures with integer labels by using the nearest neighbor resampling for training and evaluation, as shown in Table I. Last but not least, all the preprocessed images are aligned with a brain atlas image with affine transformation by using advanced normalization tools (ANTs) [43], which is open-source software for efficient medical image registration. After finishing all the preprocessing steps, the obtained affine-aligned images (moving images) could be registered to the brain atlas (fixed image) and segmented with our proposed model to demonstrate its effectiveness.

B. Evaluation Metrics

The segmentation performance is evaluated by the Dice score [70], as shown in (12). It provides the volume overlap

of the anatomical structures of segmented brain images

$$\text{Dice}_{\text{seg}} = 2 \cdot \frac{|\mathcal{S}_m \cap S_m|}{|\mathcal{S}_m| + |S_m|} \tag{12}$$

where $\mathcal{S}_m$ means the labeled anatomical segmentation of a moving brain image and S_m represents the corresponding ground truth.

Likewise, the Dice score can evaluate the registration performance [70], as follows:

$$\text{Dice}_{\text{reg}} = 2 \cdot \frac{\left|(S_m \circ \boldsymbol{\phi}) \cap S_f\right|}{|(S_m \circ \boldsymbol{\phi})| + \left|S_f\right|} \tag{13}$$

where S_m means the anatomical segmentation of a moving brain image, the counterpart S_f represents the corresponding anatomical structure of a fixed brain image, and $\boldsymbol{\phi}$ stands for the deformable registration field. When the Dice score equals 1, the corresponding anatomical structures are aligned very well after deformable registration. In contrast, the Dice score of 0 means that there is no overlap.

In addition, the diffeomorphism property should also be considered because image folding is implausible in medical image registration, and topological preservation should be encouraged. Hence, the Jacobian matrix $J_{\boldsymbol{\phi}}(p) = \nabla\boldsymbol{\phi}(p)$, which encodes the local shearing, stretching, and rotating information of the deformation field, is applied to evaluate the local property of DVF $\boldsymbol{\phi}$ around voxel p, as shown in (14).

The quantitative evaluation of diffeomorphism can be obtained by calculating the determinant of the Jacobian matrix, namely, $\det(J_{\boldsymbol{\phi}}(p))$. To be more specific, $\det(J_{\boldsymbol{\phi}}(p)) = 1$ means that no voxel deformation occurs. In addition, $\det(J_{\boldsymbol{\phi}}(p)) > 1$ indicates expansion, while $0 < \det(J_{\boldsymbol{\phi}}(p)) < 1$ indicates shrinkage taking place near the voxel p. However, $\det(J_{\boldsymbol{\phi}}(p)) \leq 0$ means that image folding occurs, which is counted to evaluate the quality of generated DVF

$$J_{\boldsymbol{\phi}}(p) = \nabla\boldsymbol{\phi}(p) = \begin{bmatrix} \frac{\partial\boldsymbol{\phi}_1}{\partial p_1} & \frac{\partial\boldsymbol{\phi}_1}{\partial p_2} & \frac{\partial\boldsymbol{\phi}_1}{\partial p_3} \\ \frac{\partial\boldsymbol{\phi}_2}{\partial p_1} & \frac{\partial\boldsymbol{\phi}_2}{\partial p_2} & \frac{\partial\boldsymbol{\phi}_2}{\partial p_3} \\ \frac{\partial\boldsymbol{\phi}_3}{\partial p_1} & \frac{\partial\boldsymbol{\phi}_3}{\partial p_2} & \frac{\partial\boldsymbol{\phi}_3}{\partial p_3} \end{bmatrix}. \tag{14}$$

C. Baseline Methods

To test and analyze the segmentation performance of our joint learning framework RSegNet, a 3-D U-Net structure-based subnetwork Seg-SubNet $\mathcal{S}_\vartheta(m)$ is exploited as the baseline to analyze the impact from the registration branch. In addition, RSegNet-f (with only forward consistency $\mathcal{L}_{\text{forward-cons}}$ compared with RSegNet) is listed for comparison.

Similarly, to assess the registration performance, comparisons to other well-performed methods are provided. The baseline methods that we selected here are two top-performing traditional methods, symmetric normalization (SyN) [34] and NiftiReg (CC) [44], and two deep learning-based methods, VM-diff [50], which is also exploited as the Reg-SubNet in our framework, and VM-diff-ss, which is semi-supervised VM-diff with auxiliary segmentation.

The SyN algorithm is implemented by using the ANT package with the fine-tuned parameter setting. The step size and the Gaussian smoothing are set to 0.25 and (9, 0.2), respectively, and 201 iterations are performed for three scales. NiftiReg is also an efficient open-source registration tool whose parameters are as follows: grid spacing of 5 (along x-, y-, and z-axes in millimeter, respectively), maximum iteration number of 1000, and standard deviation of the Gaussian kernel of 5. In addition, the parameter setting for VM-diff is given as follows: learning rate of 1×10^{-4}, regularization parameter of 1, batch size of 1, step per epoch of 100, and the number of epochs of 800. The parameter setting for VM-diff-ss is the same as that of VM-diff, with an additional segmentation hyperparameter of 0.2.

Furthermore, DeepAtlas provides a general solution to cope with joint training of registration and segmentation when many images are with few manual segmentations. The performances of the segmentation and the weakly-supervised registration trained with all the segmentation labels provide its upper bound. Thus, when comparing RSegNet with Seg-SubNet and VM-diff-ss, we compare RSegNet with DeepAtlas's upper performance bound.

D. Implementation and Comparison

Our proposed RSegNet is implemented by Keras with a Tensorflow backend, and it is implemented on Nvidia TITAN RTX GPU with 24 GB memory. The learning rate, the batch size, and the step of the epoch are set to 1×10^{-4}, 1, and 100, respectively. Before we start the joint training of RSegNet, we can pretrain the two subnetworks separately for sufficient epochs. Here, we set the number of pretraining epochs to 600 and the number of joint training epochs to 200. It requires approximately 50 h to train the model, while the prediction in the test stage is quite fast, with less than 1 s. Particularly, the hyperparameter analysis in our article is much more challenging, which involves three hyperparameters α, β_1, and β_2 corresponding to different loss terms, respectively. Thus, each aspect needs to be weighed carefully, when choosing the hyperparameters. However, sweeping all the hyperparameters with grid search is quite time-consuming in our application. Considering the similarity of the loss items corresponding to the hyperparameters α, β_1, and β_2, which are all designed with a multiclass Dice metric, we initially assume $\alpha = \beta_1 = \beta_2$ when performing hyperparameter tuning on the validation set. Fig. 4 shows the average Dice scores of the validation set for different values of the three hyperparameters α, β_1, and β_2. The results vary smoothly over a large range of the hyperparameter values, illustrating that our model is robust to the choice of α, β_1, and β_2, and we find the relatively good Dice results for both registration and segmentation exist around $\alpha = \beta_1 = \beta_2 = 0.1$, namely, the point $-1 = \log_{10} 0.1$, as shown in Fig. 4. Then, we constantly fine-tuned the hyperparameters at the close range of this point via empirical analysis and extensive search and found a group of values with the best performance on the validation set, namely, $\alpha = 0.05$, $\beta_1 = 0.1$, and $\beta_2 = 0.2$.

To evaluate the registration and segmentation performance of RSegNet, we test it on four public datasets, CUMC12,

TABLE II
SUMMARY OF REGISTRATION RESULTS ON TOTAL DATASET (MEAN/STD DEV)

Method	Averaged Dice	GPU sec	CPU sec	$\det(J_\phi(p)) < 0$
Affine	0.660 (0.166)	~	~	~
ANTs(SyN)	0.750 (0.141)	~	2143 (371)	9765 (2821)
NiftyReg	0.752 (0.135)	~	214 (54)	66895 (913)
Reg-SubNet (VM-diff)	0.756 (0.122)	0.65 (0.010)	~	5.0 (5.7)
RSegNet	0.764 (0.121)	0.89 (0.008)	~	6.0 (6.9)

TABLE III
SUMMARY OF REGISTRATION RESULTS ON THE TOTAL DATASET (HALF ROIS OBSERVED) (MEAN/STD DEV)

Method	Cerebellum-Cortex	Caudate	Putamen	Pallidum	3th-Ventricle	4th-Ventricle	Amygdala	CSF
Reg-SubNet	0.843 (0.032)	0.781 (0.036)	0.798 (0.038)	0.737 (0.054)	0.782 (0.025)	0.778 (0.030)	0.747 (0.029)	0.639 (0.046)
RSegNet	**0.848** (0.032)	**0.791** (0.031)	**0.817** (0.023)	**0.771** (0.039)	**0.788** (0.031)	**0.781** (0.032)	**0.761** (0.035)	**0.656** (0.050)

TABLE IV
SUMMARY OF SEGMENTATION RESULTS ON THE TOTAL DATASET (HALF ROIS OBSERVED) (MEAN/STD DEV)

Method	Cerebellum-Cortex	Caudate	Putamen	Pallidum	3th-Ventricle	4th-Ventricle	Amygdala	CSF
Seg-SubNet	0.823 (0.035)	0.734 (0.053)	0.810 (0.044)	0.772 (0.058)	0.646 (0.098)	0.640 (0.106)	0.742 (0.054)	0.625 (0.071)
RSegNet-f	0.853 (0.041)	0.806 (0.030)	0.838 (0.022)	0.785 (0.045)	0.750 (0.060)	0.776 (0.047)	0.769 (0.040)	0.677 (0.054)
RSegNet	**0.866** (0.031)	**0.821** (0.023)	**0.846** (0.022)	**0.800** (0.042)	**0.758** (0.057)	**0.776** (0.054)	**0.795** (0.037)	**0.693** (0.048)

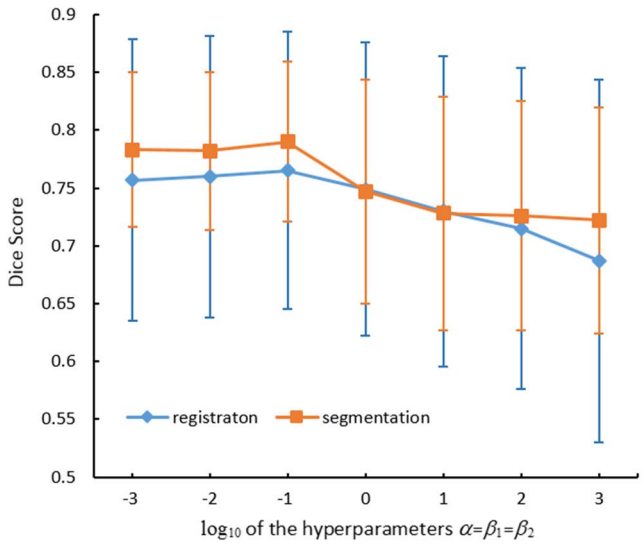

Fig. 4. Dice score of validation data for RSegNet with varied hyperparameters $\alpha = \beta_1 = \beta_2$.

IBSR18, LPBA40, and MGH10, and compare the corresponding results with the baseline methods mentioned above, respectively. In light of the GPU's memory limitation that we used for implementing our algorithm, all the ROIs listed in Table I cannot be fully exploited for the segmentation task in our framework. If we treat the corresponding anatomical structures in the left and right brain hemispheres as one region, eight ROIs (half ROIs) in maximum can be processed. Cerebellum-cortex, caudate, putamen, pallidum, 3th-ventricle, 4th-ventricle, amygdala, and CSF are selected as observed ROIs for the experiments.

As shown in Table II, four different deformable registration algorithms are compared regarding the Dice score, the computational time, and the number of voxels with nonpositive Jacobian determinants. Besides, the affine registration is also listed as a reference. We find that the deep learning-based method VM-diff achieves comparable registration accuracy compared with traditional methods, and the diffeomorphic constraint leads to a smooth and reasonable deformation field particularly. Furthermore, RSegNet performs significantly better than VM-diff with the highest average Dice score (0.764). In addition, RSegNet and the other learning-based methods require less than 1 s running on GPU, which is over 100 times faster than ANTs (SyN) and NiftyReg running on CPU.

Fig. 5 illustrates the Dice scores of all the different anatomical structures listed in Table I for four registration methods: ANTs (SyN), NiftyReg, Reg-SubNet (VM-diff), and RSegNet. Dice scores of the corresponding anatomical structures from both the left and the right brain hemispheres are averaged into one score only. We find that our RSegNet performs much better over the eight selected RoIs with the help of the auxiliary information provided by the segmentation branch $\mathcal{S}_\vartheta(m)$ and the consistency supervision and achieves comparable Dice results for all the other structures. More details about the eight selected ROIs are shown in Table III. We find significant improvements in some specific ROIs in terms of Dice score, such as putamen (+1.9%), pallidum (+3.4%), amygdala (+1.4%), and CSF (+1.7%).

Table IV presents the segmentation results of Seg-SubNet $\mathcal{S}_\vartheta(m)$, RSegNet-f, and RSegNet over the eight selected ROIs

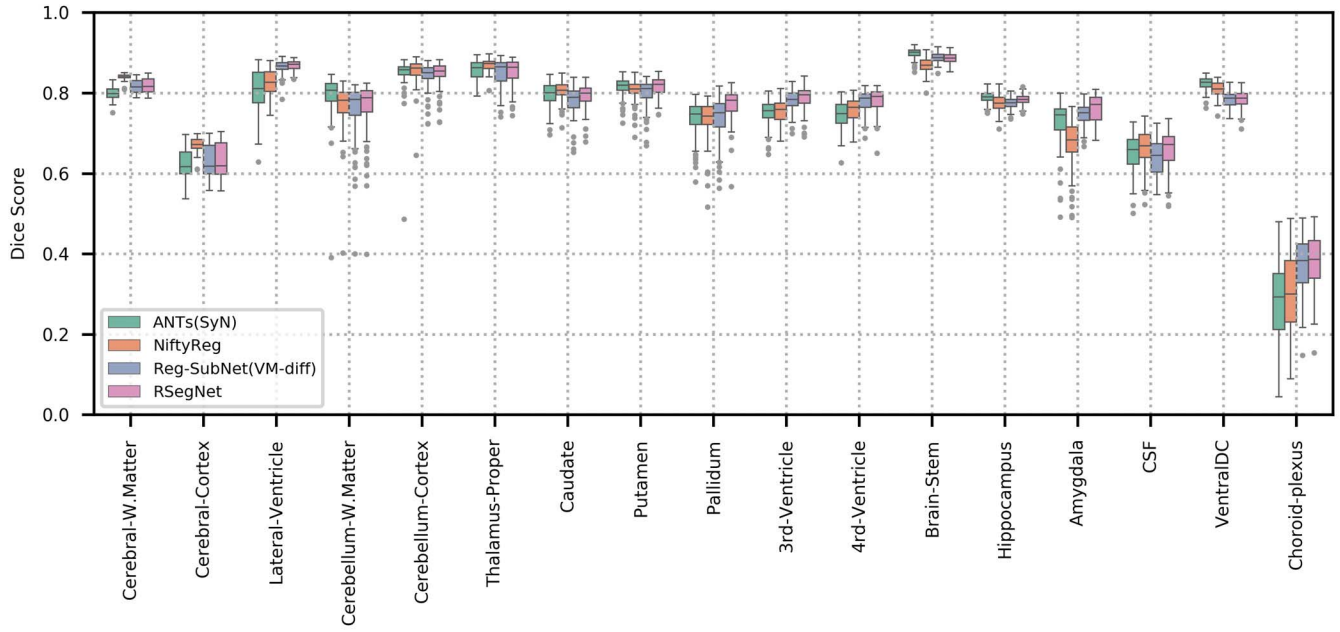

Fig. 5. Boxplot of Dice scores of different anatomical structures for four registration methods: ANTs (SyN), NiftyReg, Reg-SubNet (VM-diff), and RSegNet. Dice scores of the corresponding anatomical structures from the left and the right brain hemispheres are averaged into one score.

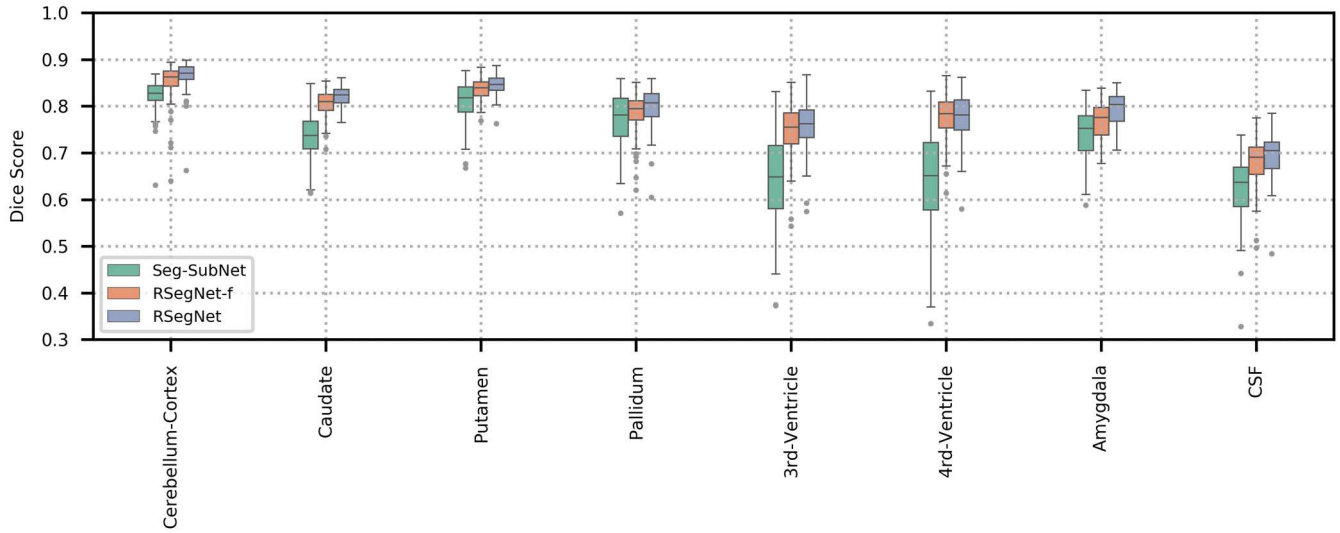

Fig. 6. Boxplot of Dice scores of various anatomical structures for three segmentation methods: Seg-SubNet, RSegNet-f, and RSegNet. Dice scores of the corresponding anatomical structures from the left and the right brain hemispheres are averaged into one score.

on the total dataset. The averaged Dice scores of RsegNet-f are all higher than those of Seg-SubNet. It means that RSegNet-f can effectively improve the segmentation accuracy with the registration branch's help through forward consistency. Moreover, RSegNet can further improve the segmentation performance against Seg-SubNet with dual-consistency, especially for caudate (+8.7%), 3th-ventricle (+11.2%), 4th-ventricle (+13.6%), amygdala (+5.3%), and CSF (+6.8%). Likewise, a boxplot of the segmentation results is displayed in Fig. 6, providing a more intuitive and straightforward comparison between these two algorithms.

To further analyze the accuracy and generalizability of our proposed RSegNet, we test it on each dataset separately. Table V presents the Dice score and a count of voxels with a nonpositive Jacobian determinant on each dataset for three registration methods, namely, Reg-SubNet (VM-diff), RSegNet, and VM-diff-ss. RSegNet can improve the means of the Dice scores on all the datasets. Moreover, we further evaluate the registration performance on the eight selected ROIs, as shown in Table VI. We observe that our RSegNet outperforms Reg-SubNet on all the subdatasets with higher mean values and can increase the Dice score by 1.4% (=0.777-0.763) on total test datasets. Besides, the performance of Reg-SubNet is very close to that of VM-diff-ss, which is the upper performance bound of DeepAtlas. Similarly, RSegNet provides a consistent increase in segmentation Dice scores on

TABLE V
REGISTRATION RESULTS ON FOUR PUBLIC DATASETS (ON ALL ROIS) (MEAN/STD DEV)

Method	CUMC12		IBSR18		LPBA40		MGH10		Total Test Sets	
	Dice	$\det(J_\phi) < 0$	Dice	$\det(J_\phi) < 0$	Dice	$\det(J_\phi) < 0$	Dice	$\det(J_\phi) < 0$	Dice	$\det(J_\phi) < 0$
Reg-SubNet	0.779 (0.114)	4.1 (3.4)	0.757 (0.111)	2.2 (3.3)	0.753 (0.130)	7.7 (6.7)	0.739 (0.110)	2.1 (2.6)	0.756 (0.122)	5.0 (5.7)
RSegNet	0.783 (0.116)	4.3 (3.9)	0.765 (0.112)	2.6 (3.3)	0.761 (0.128)	8.9 (8.1)	**0.756** (0.111)	2.4 (3.3)	0.764 (0.121)	6.0 (6.9)
VM-diff-ss	**0.786** (0.114)	2.8 (2.0)	**0.766** (0.112)	3.0 (4.7)	**0.763** (0.129)	5.8 (6.5)	0.752 (0.110)	1.4 (1.7)	**0.766** (0.122)	4.2 (5.5)

TABLE VI
REGISTRATION RESULTS ON FOUR PUBLIC DATASETS (ON EIGHT SELECTED ROIS) (MEAN/STD DEV)

Method	CUMC12	IBSR18	LPBA40	MGH10	Total Test Sets
Reg-SubNet	0.788 (0.057)	0.763 (0.061)	0.763 (0.071)	0.733 (0.061)	0.763 (0.067)
RSegNet	0.795 (0.060)	0.776 (0.063)	0.777 (0.063)	0.755 (0.057)	0.777 (0.063)
VM-diff-ss	**0.803** (0.053)	**0.782** (0.061)	**0.782** (0.061)	**0.758 (0.055)**	**0.782 (0.060)**

TABLE VII
SEGMENTATION RESULTS ON FOUR PUBLIC DATASETS (ON EIGHT SELECTED ROIS) (MEAN/STD DEV)

Method	CUMC12	IBSR18	LPBA40	MGH10	Total Sets
Seg-SubNet	0.721 (0.101)	0.725 (0.107)	0.721 (0.097)	0.738 (0.102)	0.724 (0.101)
RSegNet-f	0.806 (0.055)	0.787 (0.068)	0.773 (0.068)	0.777 (0.067)	0.782 (0.067)
RSegNet	**0.809** (0.056)	**0.794** (0.071)	**0.791** (0.063)	**0.791** (0.067)	**0.794** (0.065)

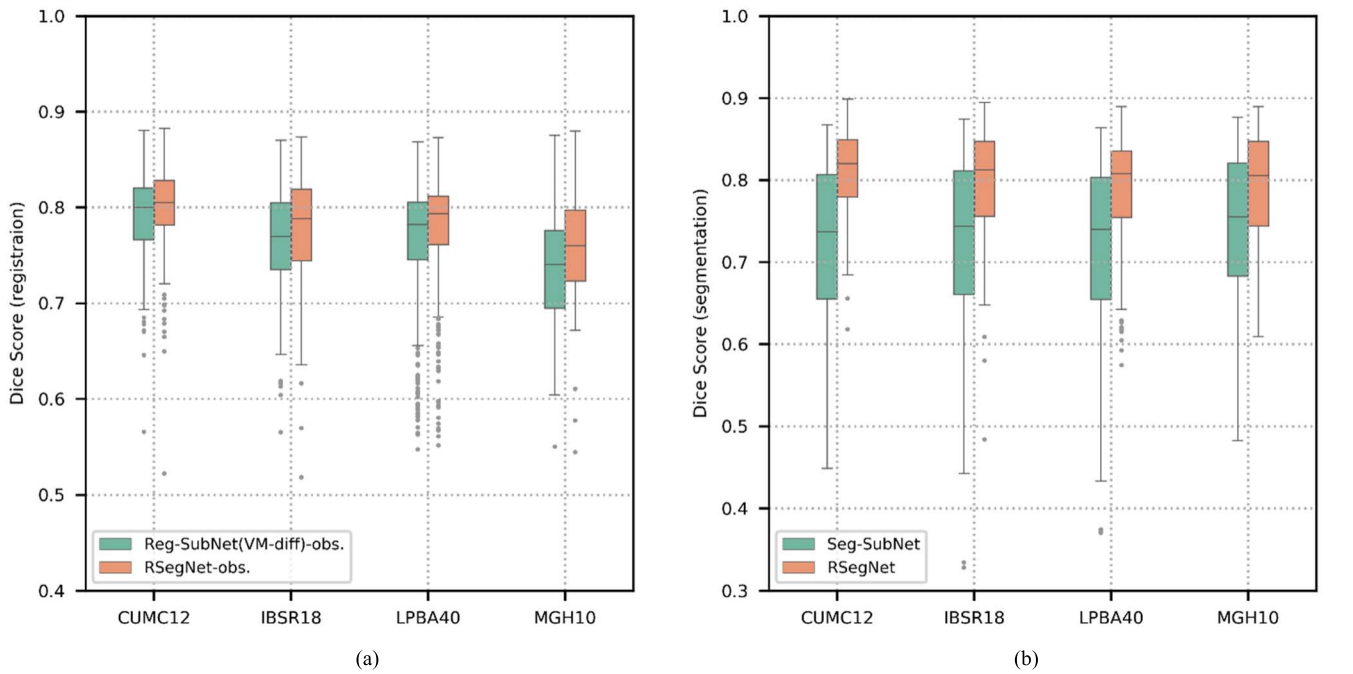

Fig. 7. Boxplot of Dice scores of registration and segmentation results on four public datasets. (a) Comparison between two registration methods: Reg-SubNet (VM-diff) and our RSegNet. (b) Comparison between two segmentation methods: Seg-SubNet and our RSegNet.

all the subdatasets compared with Seg-SubNet and has 7.0% (=0.794-0.724) improvement on entire datasets, as shown in Table VII. For better visualization, boxplots of the Dice scores for both registration and segmentation results in terms of eight observed ROIs on four datasets are shown in Fig. 7, displaying the improvements of our joint learning framework RSegNet on both tasks.

Furthermore, to provide a more intuitive visualization of the improvements, some examples are shown in this article. Fig. 8(a) illustrates registration examples on the IBSR18 dataset with four different methods: ANTs (SyN), NiftyReg, Reg-SubNet (VM-diff), and RSegNet that perform deformable registration between the moving image (Column 1) and the fixed image (Column 6). It shows that RSegNet

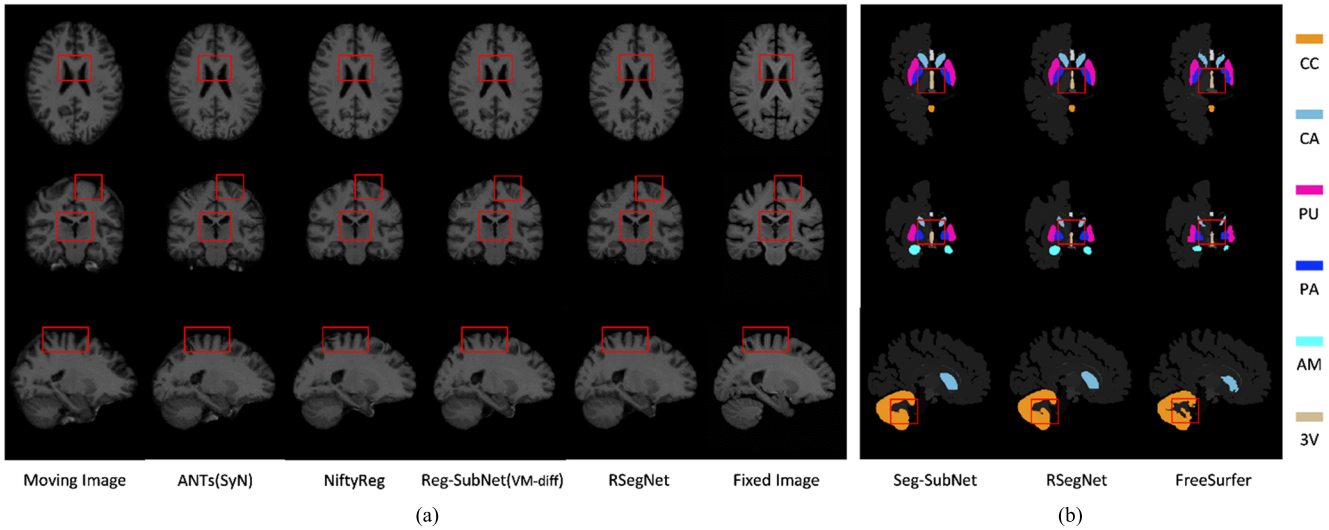

Fig. 8. (a) Registration examples on the IBSR18 dataset with four different methods, ANTs (SyN), NiftyReg, Reg-SubNet (VM-diff), and RSegNet, which register the moving image (Column 1) to the fixed image (Column 6). (b) Segmentation examples on the IBSR18 dataset with three different methods: Seg-SubNet, RSegNet, and FreeSurfer. Different colors are exploited for illustration of various anatomical structures, namely, cerebellum-cortex (CC), caudate (CA), putamen (PU), pallidum (PA), amygdala (AM), and 3rd-ventricle (3V).

can predict more accurate boundaries of various anatomical structures against Reg-SubNet (VM-diff), such as the areas marked by the red boxes, namely, caudate and lateral ventricle (first row), CSF, 3th-ventricle, cerebral white matter (second row), and cerebral cortex (last row).

On top of that, Fig. 8(b) displays segmentation examples on the IBSR18 dataset with three different methods: Seg-SubNet, RSegNet, and FreeSurfer (ground truth). We can observe that RSegNet can produce results much more similar to FreeSurfer than Seg-SubNet in terms of the 3th-ventricle and cerebellum-cortex highlighted by red boxes. Specifically, RSegNet cares more about the anatomical boundaries and outputs more accurate segmentation, while Seg-SubNet ignores these boundary details sometimes.

Based on all the experimental result analysis, we can find that our joint learning framework improves both registration and segmentation performances over separately learned subtask networks and leads to state-of-the-art accuracy and generalizability for the datasets, demonstrating the feasibility and the effectiveness of consistency supervision connecting registration and segmentation tasks.

Particularly, our experiments have trained our model with adequate training data acquired from various datasets to meet the data generality and used it for prediction or inferring on the same type of image data with a similar probability distribution. Besides, the model can rarely work on the raw data, so all the training and test data should undergo the same preprocessing procedure to output a satisfying unified data format, as aforementioned in this article. If the distributions of new data slightly deviate from those of the training data, retraining based on the trained model could quickly adapt to the target data domain and relieve the laborious training process. Suppose that the incoming new data are significantly different from the original training data regarding feature distributions, image modalities, and various anatomies (e.g., different ROI selections in our experiments). In that case, the framework should be retrained from scratch to avoid degraded or terrible predictive performances. In general, it is imperative to monitor data distribution and model performance with reasonable solutions or workflows to enable model retraining at the right time.

VI. Conclusion and Discussion

This article has proposed an end-to-end joint learning framework for simultaneous deformable registration and segmentation, named RsegNet, which establishes a principled and practical connection between these two tasks and shows significant improvements over the separately learned networks.

Our method leverages a CNN to realize multitask learning for fast runtime, reducing the computation time from minutes to under a second on a GPU compared with traditional methods.

By introducing a consistency supervision loss, the diffeomorphic registration branch and the segmentation branch can be naturally incorporated into one joint framework during the training procedure and boost each other's performances at the test stage. The probabilistic diffeomorphic registration branch in RSegNet could benefit from the auxiliary segmentation information available from the segmentation branch, which leads to improved registration results with enhanced consistency anatomically. Likewise, reasonable data augmentation based on the registration network with well-behaved diffeomorphic guarantees could also improve the segmentation performance.

Experiments on the human brain 3-D MR images have been carried out to demonstrate our approach's feasibility, and extensive analyses of the algorithm are provided. We tested its performance on four public datasets, including CUMC12, IBSR18, MGH10, and LPBA40, which shows that our method successfully yields concurrent improvements on both segmentation and registration tasks, in contrast to independently trained networks. Specifically, the Dice score of registration

is increased by almost 1% over all the anatomical structures listed in Table I and 1.4% over the eight selected ROIs with our method. Moreover, the segmentation accuracy measured by the Dice score has also been improved on each test dataset and increased by 7.0% on average. All the experimental results demonstrate the feasibility and effectiveness of our framework by exploiting each subnetwork characteristic jointly.

To sum up, our joint learning neural network, RSegNet, is a general framework that is fast, accurate, robust, and could be widely used in other image modalities or anatomies for both registration and segmentation. In the future, we will apply our model to other medical applications to further demonstrate our algorithm's generalizability.

REFERENCES

[1] A. Dalca, M. Rakic, J. Guttag, and M. Sabuncu, "Learning conditional deformable templates with convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 804–816.

[2] L. Qiu, C. Li, and H. Ren, "Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural network," *Healthcare Technol. Lett.*, vol. 6, no. 6, pp. 159–164, Dec. 2019.

[3] L. Qiu and H. Ren, "Endoscope navigation and 3D reconstruction of oral cavity by visual SLAM with mitigated data scarcity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2197–2204.

[4] Y. Yuan *et al.*, "Prostate segmentation with encoder-decoder densely connected convolutional network (ED-DenseNet)," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 434–437.

[5] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.

[6] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2933–2941.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[9] C. Baillard, P. Hellier, and C. Barillot, "Segmentation of brain 3D MR images using level sets and dense registration," *Med. Image Anal.*, vol. 5, no. 3, pp. 185–194, Sep. 2001.

[10] Y. Wu *et al.*, "Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI," *NeuroImage*, vol. 32, no. 3, pp. 1205–1215, Sep. 2006.

[11] J. M. Lötjönen *et al.*, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, Feb. 2010.

[12] X. Han and B. Fischl, "Atlas renormalization for improved brain MR image segmentation across scanner platforms," *IEEE Trans. Med. Imag.*, vol. 26, no. 4, pp. 479–486, Apr. 2007.

[13] B. C. Vemuri, J. Ye, Y. Chen, and C. M. Leonard, "Image registration via level-set motion: Applications to atlas-based segmentation," *Med. Image Anal.*, vol. 7, no. 1, pp. 1–20, Mar. 2003.

[14] V. Zagrodsky, V. Walimbe, C. R. Castro-Pareja, J. Xin Qin, J.-M. Song, and R. Shekhar, "Registration-assisted segmentation of real-time 3-D echocardiographic data using deformable models," *IEEE Trans. Med. Imag.*, vol. 24, no. 9, pp. 1089–1099, Sep. 2005.

[15] A. Yezzi, L. Zöllei, and T. Kapur, "A variational framework for integrating segmentation and registration through active contours," *Med. Image Anal.*, vol. 7, no. 2, pp. 171–185, Jun. 2003.

[16] A. Tsai *et al.*, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Trans. Med. Imag.*, vol. 22, no. 2, pp. 137–154, Feb. 2003.

[17] A. Tsai, W. Wells, C. Tempany, E. Grimson, and A. Willsky, "Mutual information in coupled multi-shape model for medical image segmentation," *Med. Image Anal.*, vol. 8, no. 4, pp. 429–445, Dec. 2004.

[18] Y. Yuan, D. Li, and M. Q.-H. Meng, "Automatic polyp detection via a novel unified bottom-up and top-down saliency approach," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 4, pp. 1250–1260, Jul. 2018.

[19] Y. Yuan, B. Li, and M. Q.-H. Meng, "WCE abnormality detection based on saliency and adaptive locality-constrained linear coding," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 1, pp. 149–159, Jan. 2017.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[21] Y. Gordienko *et al.*, "Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer," in *Proc. Int. Conf. Comput. Sci., Eng. Educ. Appl.* Cham, Switzerland: Springer, 2018, pp. 638–647.

[22] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: http://arxiv.org/abs/1804.03999

[23] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[24] H. R. Roth *et al.*, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2014, pp. 520–527.

[25] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.

[26] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (DV)*, Oct. 2016, pp. 565–571.

[27] A. V. Dalca, E. Yu, P. Golland, B. Fischl, M. R. Sabuncu, and J. E. Iglesias, "Unsupervised deep learning for Bayesian brain MRI segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 356–365.

[28] P. Hellier, J. Ashburner, I. Corouge, C. Barillot, and K. J. Friston, "Inter-subject registration of functional and anatomical data using SPM," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2002, pp. 590–597.

[29] B. A. Ardekani, S. Guckemus, A. Bachman, M. J. Hoptman, M. Wojtaszek, and J. Nierenberg, "Quantitative comparison of algorithms for inter-subject registration of 3D volumetric brain MRI scans," *J. Neurosci. Methods*, vol. 142, no. 1, pp. 67–76, Mar. 2005.

[30] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2000, pp. 452–461.

[31] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[32] W. M. Wells, III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Med. Image Anal.*, vol. 1, no. 1, pp. 35–51, 1996.

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

[34] B. Avants, C. Epstein, M. Grossman, and J. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, vol. 12, no. 1, pp. 26–41, Feb. 2008.

[35] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, Feb. 2005.

[36] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration—A deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, Sep. 2017.

[37] H. Sokooti, B. de Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3D convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 232–239.

[38] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "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*. Cham, Switzerland: Springer, 2017, pp. 204–212.

[39] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 196–205, Jan. 2010.
[40] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image Anal.*, vol. 52, pp. 128–143, Feb. 2019.
[41] Y. Hu *et al.*, "Weakly-supervised convolutional neural networks for multimodal image registration," *Med. Image Anal.*, vol. 49, pp. 1–13, Oct. 2018.
[42] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "VoxelMorph: A learning framework for deformable medical image registration," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1788–1800, Aug. 2019.
[43] 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, Feb. 2011.
[44] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.
[45] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi, "Deformable medical image registration using generative adversarial networks," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1449–1453.
[46] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen, "Adversarial similarity network for evaluating image alignment in deep learning based registration," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2018, pp. 739–746.
[47] L. Duan *et al.*, "Adversarial learning for deformable registration of brain MR image using a multi-scale fully convolutional network," *Biomed. Signal Process. Control*, vol. 53, Aug. 2019, Art. no. 101562.
[48] Y. Hu *et al.*, "Adversarial deformation regularization for training image registration neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 774–782.
[49] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[50] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Med. Image Anal.*, vol. 57, pp. 226–236, Oct. 2019.
[51] Y. Hu *et al.*, "Label-driven weakly-supervised learning for multimodal deformarle image registration," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1070–1074.
[52] J. Ashburner and K. J. Friston, "Unified segmentation," *NeuroImage*, vol. 26, no. 3, pp. 839–851, Jul. 2005.
[53] K. M. Pohl, J. Fisher, W. E. L. Grimson, R. Kikinis, and W. M. Wells, "A Bayesian model for joint segmentation and registration," *NeuroImage*, vol. 31, no. 1, pp. 228–239, May 2006.
[54] V. Duay, X. Bresson, J. S. Castro, C. Pollo, M. B. Cuadra, and J.-P. Thiran, "An active contour-based atlas registration model applied to automatic subthalamic nucleus targeting on MRI: Method and validation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2008, pp. 980–988.
[55] A. Yezzi, L. Zollei, and T. Kapur, "A variational framework for joint segmentation and registration," in *Proc. IEEE Workshop Math. Methods Biomed. Image Anal. (MMBIA)*, Dec. 2001, pp. 44–51.
[56] P. P. Wyatt and J. A. Noble, "MAP MRF joint segmentation and registration of medical images," *Med. Image Anal.*, vol. 7, no. 4, pp. 539–552, Dec. 2003.
[57] C. Xiaohua, M. Brady, and D. Rueckert, "Simultaneous segmentation and registration for medical image," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2004, pp. 663–670.
[58] Z. Xu and M. Niethammer, "DeepAtlas: Joint semi-supervised learning of image registration and segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 420–429.
[59] L. Qiu and H. Ren, "U-RSNet: An unsupervised probabilistic model for joint registration and segmentation," *Neurocomputing*, vol. 450, pp. 264–274, Aug. 2021.
[60] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95–113, Oct. 2007.
[61] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-Euclidean framework for statistics on diffeomorphisms," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2006, pp. 924–931.
[62] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[63] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *J. Cognit. Neurosci.*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007.
[64] S. G. Mueller *et al.*, "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI)," *Alzheimer's Dementia*, vol. 1, no. 1, pp. 55–66, Jul. 2005.
[65] A. Di Martino *et al.*, "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, p. 659, 2014.
[66] T. Consortium, "The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience," *Frontiers Syst. Neurosci.*, vol. 6, p. 62, Sep. 2012.
[67] A. Klein *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, Jul. 2009.
[68] D. W. Shattuck *et al.*, "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, no. 3, pp. 1064–1080, Feb. 2008.
[69] B. Fischl, "FreeSurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, Aug. 2012.
[70] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.

Liang Qiu (Student Member, IEEE) received the M.S. degree in control engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2016, and the Ph.D. degree in biomedical engineering from the National University of Singapore (NUS), Singapore, in 2021.

He is currently a Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include medical robotics, vision-based surgical navigation, and medical image registration and segmentation.

Hongliang Ren (Senior Member, IEEE) received the Ph.D. degree in electronic engineering with a specialization in biomedical engineering from The Chinese University of Hong Kong, Hong Kong, in 2008.

He was a Research Fellow with the Laboratory for Computational Sensing and Robotics, the Engineering Center for Computer-Integrated Surgical Systems and Technology, the Department of Biomedical Engineering, the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA, from 2008 to 2010. In 2010, he joined the Pediatric Cardiac Biorobotics Laboratory, Department of Cardiovascular Surgery, Children's Hospital Boston, Boston, MA, USA, and the Harvard Medical School, Boston, for investigating the beating heart robotic surgery system. In 2012, he was with the Collaborative Computer Integrated Surgery Project, Surgical Innovation Institute, Children's National Medical Center, Washington, DC, USA. He is currently an Associate Professor and leading a research group on medical mechatronics with the Electronic Engineering Department, The Chinese University of Hong Kong, and the Biomedical Engineering Department, National University of Singapore, Singapore. His research interests include intelligent systems, mechatronics, signal processing, computer-integrated surgery, and dynamic positioning in medicine.