Deformable Image Registration Using a Cue-Aware Deep Regression Network

Xiaohuan Cao, Jianhua Yang, Jun Zhang, Qian Wang, Pew-Thian Yap, *Senior Member, IEEE*, and Dinggang Shen, *Fellow, IEEE*

***Abstract*—*Significance:* Analysis of modern large-scale, multicenter or diseased data requires deformable registration algorithms that can cope with data of diverse nature. *Objective:* We propose a novel deformable registration method, which is based on a cue-aware deep regression network, to deal with multiple databases with minimal parameter tuning. *Methods:* Our method learns and predicts the deformation field between a reference image and a subject image. Specifically, given a set of training images, our method learns the displacement vector associated with a pair of reference–subject patches. To achieve this, we first introduce a key-point truncated-balanced sampling strategy to facilitate accurate learning from the image database of limited size. Then, we design a cue-aware deep regression network, where we propose to employ the contextual cue, i.e., the scale-adaptive local similarity, to more apparently guide the learning process. The deep regression network is aware of the contextual cue for accurate prediction of local deformation. *Results and Conclusion*: Our experiments show that the proposed method can tackle various registration tasks on different databases, giving consistent good performance without the need of manual parameter tuning, which could be applicable to various clinical applications.**

***Index Terms*—Deformable registration, deep learning, nonlinear regression, key-points sampling.**

Manuscript received September 25, 2017; revised January 6, 2018 and March 12, 2018; accepted March 31, 2018. Date of publication April 4, 2018; date of current version August 20, 2018. This work was supported in part by the NIH under Grant CA206100 and Grant AG053867, in part by the National Key Research and Development Program of China under Grant 2017YFC0107600, in part by the National Natural Science Foundation of China under Grant 61473190, Grant 81471733, and Grant 61401271, and in part by the Science and Technology Commission of Shanghai Municipality under Grant 16511101100 and Grant 16410722400. *(Corresponding authors: Qian Wang and Dinggang Shen.)*

X. Cao is with the School of Automation, Northwestern Polytechnical University, and also with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill.

J. Yang is with the School of Automation, Northwestern Polytechnical University.

J. Zhang is with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill.

Q. Wang is with the School of Biomedical Engineering, Institute for Medical Imaging Technology, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: wang.qian@sjtu.edu.cn).

P. Yap is with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill.

D. Shen is with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea (e-mail: dgshen@med.unc.edu).

Digital Object Identifier 10.1109/TBME.2018.2822826

I. Introduction

IMAGE registration is a crucial and fundamental procedure in various medical image analysis tasks. The aim of a registration algorithm is to obtain a topology-preserving deformation field that warps and matches a subject image to a reference image space. It can establish the anatomical correspondences between a pair of images, and thus ensures image data comparability to facilitate the subsequent analysis, e.g., group comparison or longitudinal studies. Despite the plethora of existing registration methods, image registration is still an active area of research, especially in view of the additional challenges posed by *large-scale data*, *multi-center data* (i.e., the data acquired from different institutions or under different imaging protocols), or *diseased data* with significantly heterogeneous pathology. Modern deformable registration methods should be sufficiently versatile to deal with diverse imaging data. Accordingly, it requires the algorithm consistently **accurate** (to different registration tasks), **robust** (with minimal parameter tuning), and also **applicable** to different databases and clinical scenarios.

In this paper, we introduce an approach based on deep regression networks to predict the deformation field between a pair of images that may potentially pose various challenges. Deep learning techniques, such as convolutional neural network (CNN), have been widely applied in medical image analysis [1] due to its strong learning ability, such as disease diagnosis [2]–[4], image segmentation [5], [6], and image synthesis [7]. However, there are a number of challenges for deep learning to successfully model the complex mapping of the deformable registration task, and accordingly our method has the following characteristics.

First, directly learning the mapping from the image pair to their desired deformation field is complex, since the deformation field encodes the local matching association between the pair of the reference image and the moving subject image. To simplify the problem, some traditional learning-based registration methods [8], [9] establish the mapping between the moving subject image and its deformation field by referring to a common reference image. However, when the reference is changed, the model has to be retrained from scratch. To make the registration algorithms more flexible, currently, some learning-based methods [10], [11] are not limited to a specific reference, and thus are applicable to an arbitrary pair of reference and subject images. Similarly, our method also performs flexible pairwise registration, without referring to any specific reference. This is achieved

by learning the association between any arbitrary pair of 3D local patches and their deformations. We devise an approach based on key points to obtain adequate patch samples from a database of limited size, to more effectively guide the learning process. *Second*, the reference and subject within an image pair associate with different morphological spaces. This may further increase the difficulty during the learning process. While, our method explicitly learns the differences of the spaces in which the two images reside. We particularly introduce the *auxiliary contextual cue,* i.e., the local similarity map, to enhance the awareness of the deep network to improve the learning. *Third*, the deformable image registration is an ill-posed problem and the estimated local deformation may have ambiguity if the local patch cannot contain sufficient anatomical details. To address this issue, we propose a novel sampling strategy to sparsely sample representative patches from the image pair to avoid the ambiguity. In general, our proposed *patch-wise cue-aware deep regression network* is able to predict the deformations accurately and robustly for the databases of significantly different natures with minimal parameter tuning, which is needed in real clinical scenarios.

A. Related Works

Comprehensive summaries of registration methods can be found in [12]–[19]. In these papers, registration algorithms differ mostly in terms of deformable models, matching criteria, and numerical optimization. While learning-based deformable registration methods are more popular recently, in which some machine-learning techniques are often incorporated in the registration framework. In the following discussion, we categorize existing registration methods as 1) *conventional methods* and 2) *learning-based methods*.

1) Conventional Registration Methods: Conventional methods regard the deformable registration as a high-dimensional optimization problem with a typical cost function:

$$\phi = \min_{\phi} \mathcal{M}\left(I_R, \mathcal{D}\left(I_S, \phi\right)\right) + \lambda \mathcal{R}\left(\phi\right), \tag{1}$$

where the deformation field ϕ can be obtained by minimizing the dissimilarity $\mathcal{M}$ between the reference image I_R and the warped subject image $\mathcal{D}(I_S, \phi)$, with regularization $\mathcal{R}$ on the deformation field ϕ in order to avoid the unpractical deformations. $\mathcal{D}$ is the operator that warps the subject image I_S using deformation field ϕ.

Aiming to solve the optimization problem in Eq. (1), a large number of registration methods have been proposed using various similarity metrics and regularization terms. Widely-used similarity metrics include sum-of-square distance (SSD) [20]–[22], mean square distance (MSD) [23]–[25], (normalized) cross correlation (CC) [26]–[29], and (normalized) mutual information (MI) [30]–[33], *etc*. Regularity of the deformation field can be achieved by Gaussian smoothing [20], [21], [23], [24], [28], [34], minimizing the bending energy, or by utilizing a spline-based [25], [29], [35], [36] or diffeomorphic [21], [29], [34], [37] deformable model. Based on the matching criterion, the deformable registration can be divided into two categories. (1) The volumetric-transformation-based registration [20], [21], [23], [24], [27], [34], in which the voxel intensity information in the whole volume is used to guide the registration. (2) The landmark-based registration [28], [29], [35], [36], [38], in which the features or attributes are used as the morphological signatures of the landmarks to drive the local correspondence matching during the registration. Additional properties, such as symmetry [28], [29], can also be imposed.

Commonly-used methods include AIR [23], ART [26], ANTS [34], Demons [20], Diffeomorphic Demons [21], [39], SPM [40], RAMMS [36], DROP [41], CC/MI/SSD-FFD [35], and FNIRT [22], *etc*. Although comprehensive performance comparisons of some of these methods are reported in [42], [43], it is still difficult to assert the best algorithms for individual applications, especially when dealing with various databases or registration tasks.

For conventional registration methods, most of them require iterative optimization and careful parameter tuning, which depends on the nature of the data. The registration performance may decline when the reference and subject images have large appearance and anatomical variations. Thus, a robust and tuning-free registration method is essential for wide utility by non-experts for various clinical studies, which should consistently work well for different databases or registration tasks.

2) Learning-Based Registration Methods: For learning-based registration methods, different machine learning techniques are often incorporated into the registration framework. The complex deformable registration problem is often simplified by leveraging prior knowledge or the parameters that are predicted by learning. The mapping between image appearances and the deformation field can be learned by support vector regression (SVR) in [8] or sparse representation in [9]. Initial deformation field can be predicted rapidly using the learned models and then refined effectively by one of existing deformable registration methods. As the refining is much less demanding compared to estimation of the whole deformation field directly, the above methods have demonstrated improved registration performances especially for the algorithm efficiency. However, the learned models are usually associated with a common reference, which need to be retrained when the reference space changes. Therefore, the real-world applicability of these methods is strongly limited.

Techniques, such as random forest have been successfully applied for infant brain registration [44] and multi-modal image registration [45]–[47]. These are challenging tasks due to inconsistency in appearances and differences in structural geometries. Random forest is able to predict and compensate for large local deformations, simplifying the registration task and improving the registration accuracy. However, the effectiveness of this approach highly depends on the hand-engineered features, which is a crucial factor during the learning of random forest.

In contrast, deep learning conducts the learning process from the raw image without the need of feature engineering. For linear registration, CNN is used to align X-ray images in [48] and deep reinforcement learning is applied to align the CT and depth images in [49]. For deformable registration, features are learned automatically in an unsupervised manner in [50] and then incorporated into feature-based registration. Some works learn the transformation parameters in a supervised manner. Specifically, fully convolutional network (FCN) has been employed in

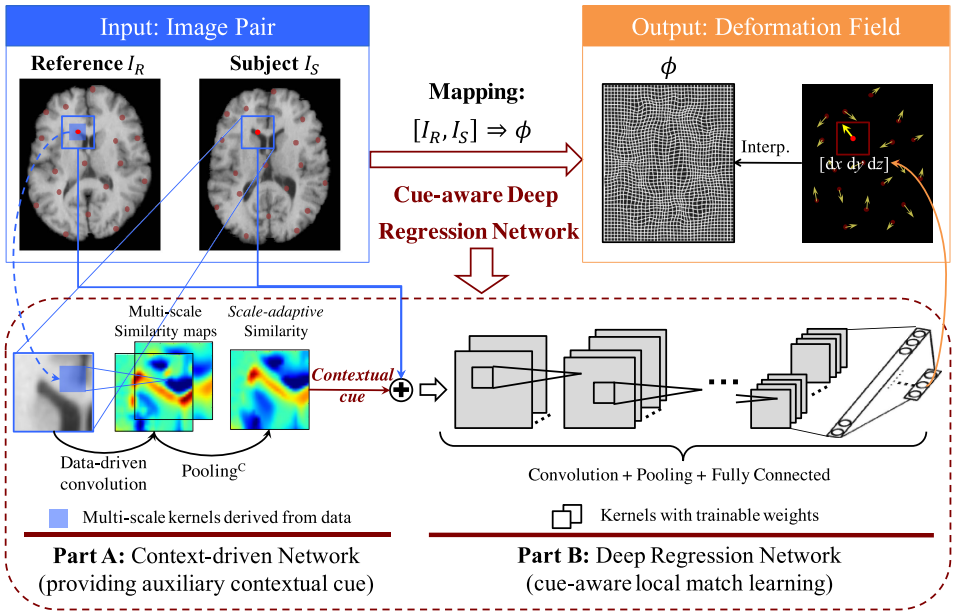

Fig. 1. Method overview: Cue-aware deep regression network for deformable image registration. The input is a pair of images and the output is the deformation field. For simplicity, examples are shown in 2D; but the actual implementation is carried out in 3D.

[10] and [51] to learn the deformation momenta and stationary velocity field (SVF), respectively. For lung CT image registration, the displacement vector is predicted by a CNN model in [11]. To align the prostates in MR images, the deformation field is estimated by a deep reinforcement learning framework [52]. Recently, unsupervised deep learning is also adopted for registration, in which the similarity metrics (e.g., normalized cross-correlation [53], [54]) are directly used to train the deep network in backpropagation.

B. Contributions of Our Work

In this work, we learn the mapping between an image pair and their deformation field. The displacement vectors will be estimated by the proposed *cue-aware deep regression network* in a patch-wise manner. In the application stage, the final deformation field for the arbitrary image pair can be effectively obtained by inputting the reference and subject images to the learned model. Our main contributions are summarized as follows.

1) To successfully establish the deep regression model for the highly non-linear mapping of deformable registration task, our network enhances generalization and robustness by introducing the *auxiliary contextual cue*, which provides robust local similarity information for participating the whole training process. This has been effectively incorporated into the whole network by *data-driven convolution* and *cross-channel pooling*.
2) To mitigate the ambiguous matching, a *key-point truncated-balanced sampling strategy* is proposed to generate a completed and well-functioning training set. It can also generate sufficient training patches from a limited image database. Under this strategy, the prediction accuracy of the deep regression model has been greatly improved.
3) To fully evaluate the proposed method, we perform comprehensive experiments on different databases with challenging registration tasks. We demonstrate that, the proposed method performs consistently well without parameter tuning even on the challenging registration tasks involving databases of diverse natures.

II. Method

A. Overview

The mapping from a reference-subject 3D image pair $[I_R, I_S] \in \Re^3$ (affinely registered) to their deformation field ϕ can be generally denoted as:

$$[I_R, I_S] \overset{M}{\Rightarrow} \phi, \tag{2}$$

where M is the mapping to be learned by the proposed deep regression network.

As shown in Fig. 1, the proposed *cue-aware deep regression network* is designed in a 3D patch-wise manner. The network (encircled by the dashed box in the bottom of Fig. 1) maps patch appearance to the corresponding displacement vector of the patch center. The mapping is learned using a training set generated by a *key-point truncated-balanced sampling strategy*. The whole learning procedure can be briefly introduced as follows. **First**, a pair of local patches with a common center location is extracted from both the reference I_R and subject I_S images. **Then**, a scale-adaptive contextual cue, which encodes multi-scale similarity information, is generated via the *context-driven network (Part A)* by performing *data-driven convolution* and *cross-channel pooling* (pooling^C). **Next**, the contextual cue, in addition to the patches, is fed into the *deep regression network (Part B)* for cue-aware local match learning. The deep regression network outputs the displacement vector associated with the

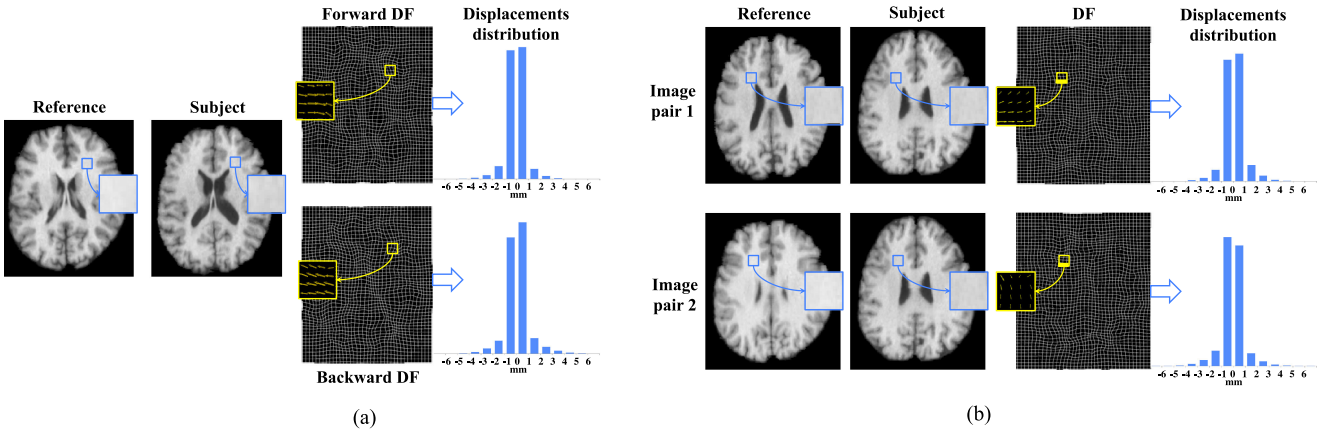

Fig. 2. Demonstration of matching ambiguity and imbalanced deformation distribution. (a) Patch pair is similar in appearance but with different forward and backward deformation fields (DF) if swapping the order of reference and subject. (b) Patch pairs are similar in appearance but associated with very different DFs.

patch center. **Finally**, in the application stage, the learned deep network is used to predict the displacement vectors at locations that sufficiently cover the whole brain. The final deformation field is obtained via spline interpolation.

B. Generating Training Data

Our deep network learns the deformation associated with 3D patch pairs. Each training sample consists of a patch pair $[p_R(\boldsymbol{u}), p_S(\boldsymbol{u})]$ (with patch size: $31 \times 31 \times 31$) extracted from the reference image I_R and the subject image I_S, and the corresponding displacement vector $\mathrm{d}\vec{\boldsymbol{u}} = [\mathrm{d}x, \mathrm{d}y, \mathrm{d}z]_{\boldsymbol{u}}$ defined in the reference space. Here, $\boldsymbol{u}$ indicates the location of the center.

Usually, random or uniform samplings are often employed to generate the training set. Fig. 2 illustrates the problems that can result from random or uniform sampling for the registration task. The *first problem* is ambiguous matching. Fig. 2(a) shows the situation where two patches with highly similar appearances have significantly different forward and backward deformation fields (DFs), if we swap the reference and the subject. Fig. 2(b) shows two similar pairs of patches that cannot provide sufficient information to differentiate DFs associated with different references. The *second problem* is imbalanced deformation distribution. As shown in Fig. 2, the deformation distribution is significantly imbalance with over 80% of displacement samples below 1mm. Conventional sampling method focus only on the *image space* and ignores the distribution in the *deformation space*. As a remedy, we propose a *key-point truncated-balanced* (**KP-TB**) sampling strategy to generate informative and representative training sets, in which the sampling regards to *not only* the **image space**, *but also* the **displacement space**.

1) KP-TB Sampling Strategy: In the *image space*, we utilize the *key-point sampling* to obtain informative patches to mitigate ambiguous matching. Obviously, brain regions with strong edges or large curvatures (e.g., ventricular boundaries, roots of sulci, crowns of gyri, and *etc*.) contain more anatomical details that can contribute to accurate matching. To extract informative patches, we generate the normalized gradient map $G(\boldsymbol{u})$ via Canny edge detector using a smoothed version of image I_R:

$$G(\boldsymbol{u}) = \frac{\sum_i \nabla_i(\boldsymbol{u})}{\|\nabla_i(\boldsymbol{u})_2\|}, i = x, y, z, \boldsymbol{u} \in I_R, \tag{3}$$

where $\nabla_i(\boldsymbol{u})$ is the gradient calculated for direction i in a 3D image space. $\|\cdot\|_2$ is the *L2*-norm used to normalize the gradients to [−1,1]. Sampling is performed based on the normalized gradient map, putting higher probability for voxels with larger gradient magnitudes.

For the *displacement space*, we introduce *truncated-balanced sampling* to obtain a training set that captures the major distribution of the displacement magnitudes. By incorporating the gradient information $G(\boldsymbol{u})$, the sampling probability is defined as:

$$P(\boldsymbol{u}) = \exp\left(\frac{-\omega}{\|\mathrm{d}\vec{\boldsymbol{u}}\|_2 \cdot |G(\boldsymbol{u})|}\right), \tag{4}$$

where $\mathrm{d}\vec{\boldsymbol{u}}_2$ is the displacement magnitude and ω is a parameter used to control the sampling probability and the sample number. $|G(\boldsymbol{u})|$ is the absolute value of $G(\boldsymbol{u})$. Apparently, an informative voxel at $\boldsymbol{u}$ (i.e., higher $|G(\boldsymbol{u})|$)) with larger displacement magnitude ($\|\mathrm{d}\vec{\boldsymbol{u}}_2\|$) is more likely to be sampled. Additionally, a truncation threshold σ is applied on the displacement magnitudes to set all $\|\mathrm{d}\vec{\boldsymbol{u}}_2\|$ to 0 when $\|\mathrm{d}\boldsymbol{u}_2\| > \sigma$:

$$\|\mathrm{d}\vec{\boldsymbol{u}}_2\| = \begin{cases} \|\mathrm{d}\vec{\boldsymbol{u}}_2\|, \|\mathrm{d}\vec{\boldsymbol{u}}_2\| \le \sigma \\ 0, \|\mathrm{d}\vec{\boldsymbol{u}}_2\| > \sigma \end{cases}. \tag{5}$$

The truncation threshold σ is applied with two reasons: (1) extremely large displacements are rare in real-world image registration problem; (2) learning a very large displacement is inefficient and requires very large input patch. Thus, we employ the truncated operation to saturate all displacements over σ, in order to guarantee the precision and generalization during the model learning. Additionally, all the displacement values are normalized to [−1,1] by dividing the truncated value with σ to adjust the subsequent network learning.

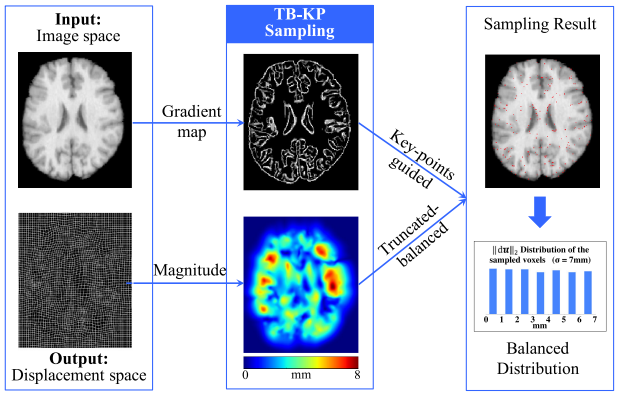

Fig. 3. Illustration of KP-TB sampling and the typically sampled locations.

Using KP-TB sampling, most samples are located at the informative regions throughout the whole brain, as shown in Fig. 3. Here we use $\sigma = 7$ mm in this paper, such that sufficient samples can be acquired while their distribution regarding the displacement magnitudes is balanced. This can well guarantee the precision and generalization during the model learning. Note that, in the application stage, the prediction of displacement will not be limited by σ. We can perform the learned model repeatedly, so that the predicted displacement can be accumulated to compose to the final accurate deformation field, with more details described in Section II-D.

C. Cue-Aware Deep Regression Network

To learn the complex appearance-to-deformation mapping, we propose to use *auxiliary contextual cue* to facilitate robust local match learning. The proposed *cue-aware deep regression network*, shown in Fig. 1, includes two main components: 1) *Part A: context-driven network*, and 2) *Part B: deep regression network*.

1) Part A: Context-Driven Network: The context-driven network is designed to provide the *axillary contextual cue* that relates two images. This cue encodes the scale-adaptive local similarity map that conveys the local correspondences from the center location in the reference to all locations in the subject patch. It can thus assist the whole network to keep aware of the local matching. The context-driven network is realized by two operations: 1) **Data-driven convolution,** which provides multi-scale similarity feature maps, and 2) **Pooling$^{\mathrm{C}}$**, which fuses multi-scale similarity maps via *cross-channel pooling*.

Data-driven Convolution. Given a patch pair $[p_R(\boldsymbol{u}_c), p_S(\boldsymbol{u}_c)]$ centered at $\boldsymbol{u}_c$, as shown in Fig. 4, the conventional similarity feature map is computed as:

$$H_S(\boldsymbol{u}_i) = \frac{\sum k_R(\boldsymbol{u}_c) \cdot k_S(\boldsymbol{u}_i)}{|k_R(\boldsymbol{u}_c)| \cdot |k_S(\boldsymbol{u}_i)|}, \boldsymbol{u}_i \epsilon p_S\ (\boldsymbol{u}_c), \tag{6}$$

where $H_S(\boldsymbol{u_i})$ is the similarity corresponding to location $\boldsymbol{u}_i$ in the subject patch p_S, and i is the location index. $k_R(\boldsymbol{u}_c)$ is a sub-region extracted from the reference patch centered at $\boldsymbol{u}_c$, and $k_S(\boldsymbol{u}_i)$ is an identical-size sub-region from the subject patch centered at $\boldsymbol{u}_i$. $|\cdot|$ denotes the $L2$-norm of the values in the sub-region. In this way, $H_S(\boldsymbol{u_i})$ represents the normalized cross-correlation and H_S is the local similarity map for each patch pair. Specifically, we implement this step as an additional data-driven convolutional layer. In this layer, the kernels are derived from the reference patch and varied according to the samples. It is thus different from the traditional convolutional layer in CNN where the kernels determine their parameters automatically during the training. As shown in Fig. 4, the output of this operation can be regarded as an external feature map with explicit contextual information of local correspondence, which guides the subsequent local match learning.

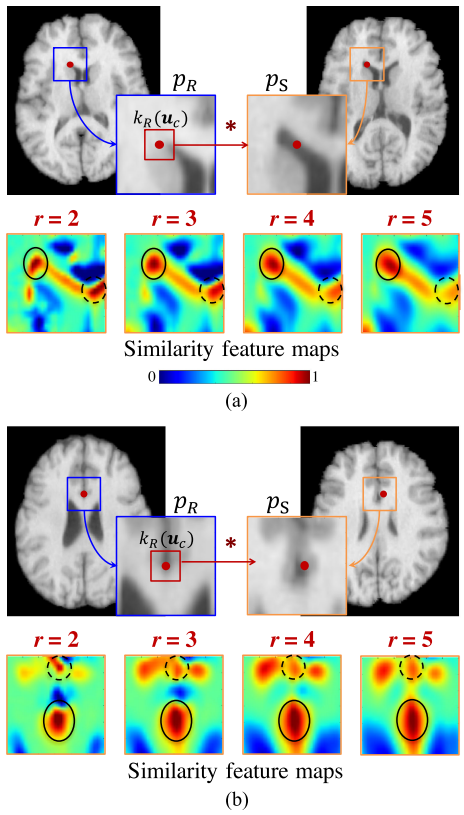

Fig. 4. Similarity maps of the two patches given by data-driven convolution using multi-scale kernels. The solid and dashed circles represent correct and incorrect guidances, respectively. r denotes the radius of the convolutional kernels. "$*$" is the data-driven convolution. A higher similarity value indicates greater correspondence with the center location.

Ideally, a contextual cue should have high responses for corresponding anatomical regions and vice-versa. Fig. 4 shows that the similarity map is sensitive to the kernel size. For example, a small kernel size results in a more distinctive similarity map, whereas a larger kernel size reduces distinctiveness but is conducive to robust matching. For both distinctiveness and robustness, we compute multiple similarity maps with multi-scale kernels as shown in Fig. 4.

Pooling$^{\mathrm{C}}$ (cross-channel pooling). It is necessary to wisely integrate the multi-scale similarity feature maps by getting rid of the redundancy from the multiple similarity maps. To tackle this issue, we introduce the pooling$^{\mathrm{C}}$ procedure to fuse the similarity feature maps by performing *pooling across channels*, as illustrated in Fig. 5. Pooling$^{\mathrm{C}}$ collapses the channels of the similarity map but retain the map size. This is different from the conventional pooling in CNN, which reduces the size of the feature maps and retain the channel number.

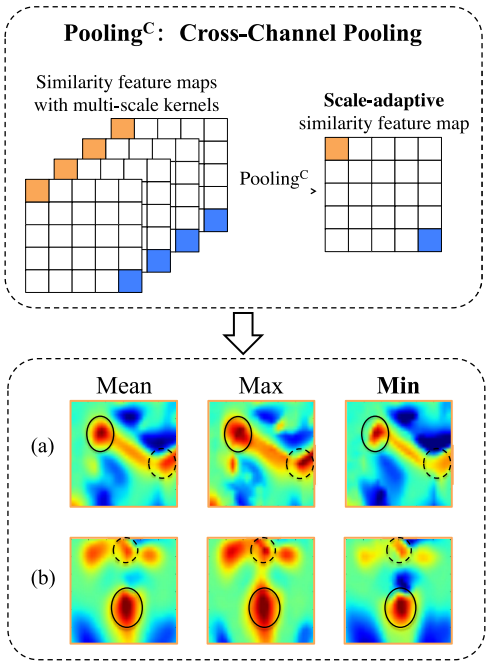

Fig. 5. Illustration of cross-channel pooling (Pooling^{C}): Nonlinear fusion of multiple similarity feature maps by getting rid of redundant information. (a) and (b) are the pooling^{C} results of the 4-scale similarity feature maps in Fig. 4.

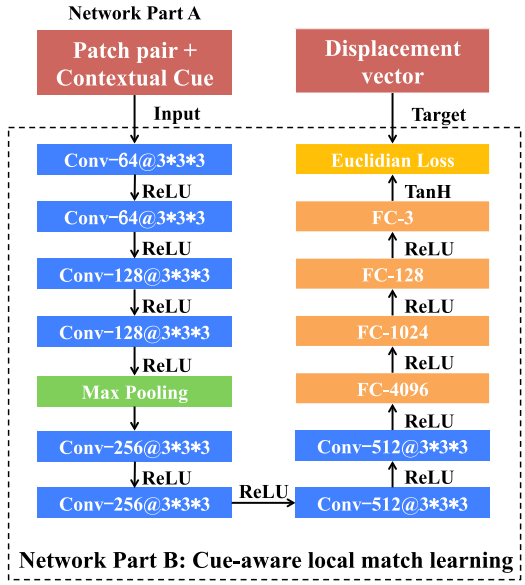

Fig. 6. The architecture of deep regression network.

The most commonly used pooling is by computing the maximum (Max) or mean (Mean). However, neither of them is appropriate for fusing the multi-scale similarity feature maps. Fig. 5 shows the pooling^{C} results of the multi-scale similarity maps in Fig. 4. Mean pooling^{C} is a trade-off between distinctiveness and robustness, whereas the result of max pooling^{C} is worse than any single similarity feature map. In this way, we introduce *minimum* (Min) operator to better preserve both distinctiveness and robustness. As shown in Fig. 5, the fusion result by min pooling^{C} is *scale-adaptive*: we can always obtain an effective fusion map, regardless of the scales used.

As a summary, the *context-driven network* can be regarded as a preparation step to provide an informative guidance, which is served as the auxiliary contextual cue, to effectively facilitate local match learning for our registration task. The data-driven convolution provides an effective way to associate the two input patches. In this specific convolution layer, the similarity feature maps are generated from the multi-scale kernels. The subsequent pooling^{C} layer fuses the multi-scale similarity feature maps, which gets rid of the redundancy within multiple feature maps yet retains their distinctiveness. By using this informative contextual cue, the awareness of local match learning can be more effectively steered in the subsequent network in *Part B: Deep Regression Network*.

2) Part B: Deep Regression Network: The deep regression network is designed to predict the displacement vectors $[\mathrm{d}\vec{\boldsymbol{u}}]$ from the patch pair $[p_R(\boldsymbol{u}), p_S(\boldsymbol{u})]$ and the guidance of the contextual cue. The detailed architecture of the network, shown in Fig. 6, consists of several convolutional layers, one pooling layer, and several fully connected layers. Specifically, each convolution layer is followed by ReLU activation to enhance the nonlinearity and modeling capability. The kernel number is doubled every two convolutional layers, starting from 64 to 512 with a fixed kernel size $3 \times 3 \times 3$. One Max pooling layer is performed to train the network efficiently. The subsequent fully connected layers include 3 layers with ReLU activations. The final fully connected layer uses TanH activation since the normalized displacement vectors are zero-centered. The loss function is the Euclidian distance. No padding operation is performed.

In order to demonstrate whether the contextual cue, i.e., the scale-adaptive similarity map, plays a positive role to enhance the awareness of the local match learning, we show the loss changing for training and validation in Fig. 7. The input patch pairs alone, without the contextual cue, are insufficient to train the deep network as shown in Fig. 7(a). While, the loss decreases consistently when including the contextual cues in Fig. 7(b) and (c). Particularly, we adopt the proposed scale-adaptive cue in Fig. 7(c), while in Fig. 7(b) we use the multi-scale contextual cue without pooling^{C}. According to more consistent and faster loss decrease achieved in Fig. 7(c), we conclude that the auxiliary contextual cue can truly enhance the awareness of the local match leaning, and the network in *Part A* contributes to *Part B*.

D. The Application Stage

In the application stage, the trained model, which is obtained via the proposed cue-aware deep regression network, can be applied directly to predict the deformation field for an unseen image pair, as summarized below.

The **first** step involves patch extraction from the image pair. It is based on the KP sampling (i.e., Eq. (3)) without TB in the reference image space. Adequate patch samples will be obtained to well cover the whole brain volume, and the patch size is exactly same with the training patch size ($31 \times 31 \times 31$).

The **second** step involves displacement prediction for each patch sample, using the trained cue-aware deep regression network to 1) generate the scale-adaptive contextual cue by the context-driven network, and then 2) estimate the displacement

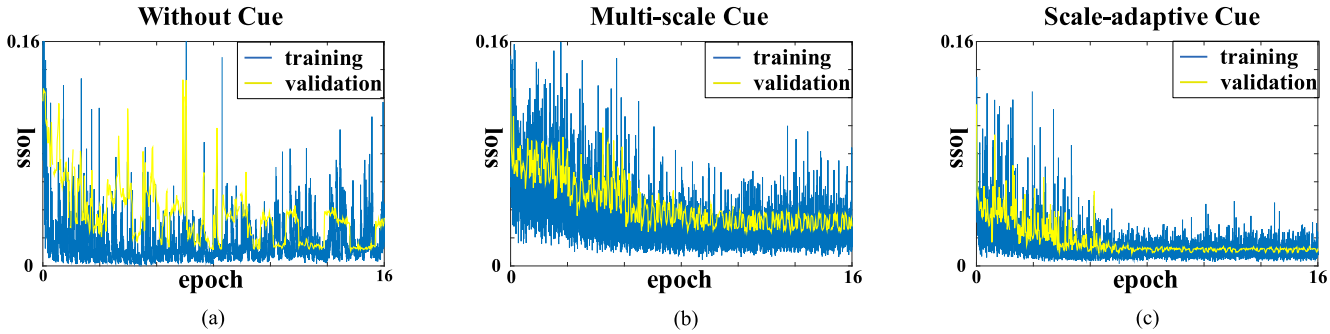

Fig. 7. Comparison of loss curves in three training scenarios: (a) without using contextual cue **(Without Cue)**, (b) using the multi-scale contextual cue without pooling^C **(Multi-scale Cue)**, and (c) using the proposed scale-adaptive contextual cue generated by pooling^C **(Scale-adaptive Cue)**.

TABLE I
DETAILED INFORMATION FOR THE THREE DATABASES: LONI LPBA40, IXI AND ADNI

Data	Image #	Train #	Test #	ROI Labels	Tissue Maps	Size & Resolution
LONI	40	25	15	54	GM, WM	220×220×220
IXI	30	0	30	83	0	
ADNI	50	0	50	0	GM, WM	$1\times1\times1\text{mm}^3$

by deep regression network. Since the output of the network is within $[-1,1]$, the real displacement magnitude is recovered by multiplication of the output with σ defined in Eq. (5).

The **third** step involves generation of the dense deformation field. Based on the displacement predictions for the adequate samples, the final deformation field can be obtained by block-wise thin-plate spline (TPS) interpolation. Details for the interpolation can be found in our early work [38].

Since we truncate the displacement magnitude during the training stage (as described in Section II-B and Eq. (5)), the largest prediction of the displacement magnitude is 7 mm (as shown in Eq. Fig. 3). Then, we can repeat the above three steps, and the final deformation field can be generated by sequentially composing the estimated deformation fields at individual iterations. The model is iteratively applied until the incremental deformation is trivial, and the registration result converges. It is worth noting that, under this strategy, although the input patch size (i.e., $31 \times 31 \times 31$), along with the truncated threshold, may limit the receptive field of the network during training, we can still accurately predict large deformations.

III. EXPERIMENTS

Three different databases, i.e., LONI LPBA40, IXI and ADNI, are used to evaluate the deformable registration performance, which cover both the young and old adult brain MR images. All the images are preprocessed using a standard pipeline, including skull stripping and resampling. The cerebellum and brain stem are also removed. After preprocessing, all data are with the same image size $220 \times 220 \times 220$ and resolution $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$. Detailed data description is provided in Table I. If not mentioned otherwise, all the image pairs used to perform the deformable registration have already been affine registered by FLIRT [55]. The gray matter (GM) and white matter (WM) tissue maps of both LONI and ADNI datasets are generated by two steps: 1) using FAST segmentation in FSL [56] to obtain a rough tissue segmentation map and then 2) performing manual correction to make them as accurate as possible.

Among the numerous registration methods, we select two state-of-the-art registration methods for comparison: 1) Demons, which is a well-known and widely-used deformable registration method. Here, we use two versions of Demons: SSD-Demons [21] and LCC-Demons [39]. These two versions use the sum of square distance (SSD) and the local correlation coefficients (LCC) as the similarity metric, respectively. As SSD-Demons is more efficient while LCC-Demons is more accurate and robust when registering images with large appearance/intensity variations, we employ SSD-Demons for registering images within one database and LCC-Demons for registering images across different databases. 2) Symmetric Normalization (SyN) [34], which has shown outstanding performance as demonstrated in [42], [43]. Dice similarity coefficients (DSC) and averaged surface distance (ASD) are used as two primary metrics to evaluate the registration performance.

Training Image Pair Generation. We randomly select 25 images from the LONI database to prepare the training data for the proposed cue-aware deep regression network. Among these 25 images, 40 image pairs are drawn randomly in the training stage. For each image pair, we generate the ground-truth deformation field in three steps. *Step 1*: We perform deformable registration on intensity images by SyN under the recommended parameter setting. *Step 2*: Then, for each image pair, we check the registration result and tune the parameters individually in order to obtain better registration quality. Here, better registration means more favorable result in visual inspection and quantitative DSC evaluation at the same time. *Step 3*: Finally, we apply Demons to align the boundaries of the manually-edited tissue segmentation maps for more accurate deformable registration. In this way, the final deformation field is generated. Note that, the manually-edited tissue segmentation maps are only used to prepare the training data.

In order to fully evaluate the performance of the proposed registration method, we carry out the experiments in three parts to gradually increase the difficulties of the registration tasks by performing the deformable registration (1) on the same database as both training and testing within LONI (Section III-A), (2) on two different databases (by training on LONI while testing

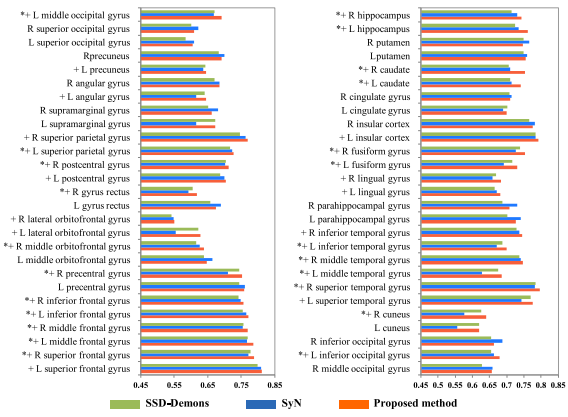

Fig. 8. Comparison of the registration results by **SSD-Demons**, **SyN** and the **proposed method**, respectively. The results are evaluated in term of DSC across the 54 ROIs in LONI LPBA40 database. "+" indicates that the proposed method outperforms the two state-of-the-art methods and "∗" means statistically significant improvement ($p < 0.05$ for paired *t*-test).

on IXI; Section III-B), and (3) across different databases (the two images within the reference-subject image pair are drawn from ADNI and LONI, respectively; Section III-C). Note that, when we use LONI data in the testing stage, we exclude the 25 training images and consider the remaining 15 images only. For the proposed method, 6% of the whole brain volume (taking no account of the background voxels) are sampled as the key points to drive the image registration in the application stage.

A. Experiments on LONI Database

We first perform the registration experiments on the LONI database by using the remaining 15 images for testing. By drawing a pair of images from 15 images, we perform registration 210 times in total for each method. In this experiment, we iterate the trained model twice, as the incremental deformations become negligible since the third iteration. The averaged results are reported as follows. Fig. 8 has shown the DSC value per ROI after registration by SSD-Demons, SyN and the proposed method. For the 54 ROIs, the proposed method has shown improved DSC values on 35 ROIs, and among them 23 ROIs are improved with statistical significance compared with the two state-of-the-art methods.

Fig. 9 has shown the comparison results of the three methods based on the tissue segmentation maps. From these results, we can observe that, in most cases, the proposed method achieves the overall best performance in terms of both DSC and ASD. Although for GM, SyN has higher accuracy, the difference is not significant. Therefore, the proposed method can at least achieve the comparable registration performance compared with the state-of-the-art methods. It is worth noting that, the proposed method only samples 6% of the whole brain image voxels to obtain the reported performance without parameter tuning. This can well demonstrate that, the trained cue-aware deep regression network is accurate and the proposed deformable registration method is well applicable.

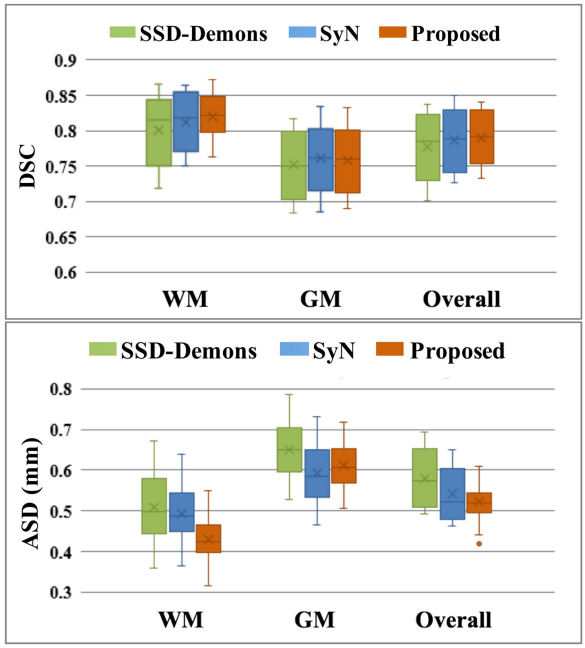

Fig. 9. DSC and ASD (mm) evaluated on the WM and GM tissue maps after performing the deformable registration by **SSD-Demons**, **SyN** and the **Proposed** method, respectively.

B. Experiments on IXI Database

To evaluate the robustness and the transferring capacity of the trained model, we directly use the learned model to perform the registration on the IXI database. For the total 30 images, we equally split into two groups that are served as the reference image group and the subject image group, respectively. In this section, we perform registration 225 times in total for each method by drawing the reference and subject from the two groups. Among 83 ROIs defined in IXI, 70 stable ROIs are used here to evaluate the registration performance. There are 13 ROIs excluded, since those ROIs are too tiny for reliable performance evaluation. In this experiment, since the new incremental deformations are almost zero in the third iteration, we also iterate the trained model two times.

The DSC values are provided in Fig. 10 for the comparison of SSD-Demons, SyN and the proposed method. From Fig. 10, we can observe that, for the 70 ROIs, the proposed method works better than both SSD-Demons and SyN in 50 ROIs. Among them, the performance of 28 ROIs are statistically significant improved, as marked by the symbols "+" and "∗" in Fig. 10. This result suggests three merits of our method. 1) The superior robustness of the proposed method. The generalization of the learned model has been well demonstrated in this experiment since we successfully apply the model on a different database and achieve promising registration performance. 2) The high accuracy of the proposed method. The performance is at least comparable with the state-of-the-art registration methods, while for most ROIs the proposed method has shown even better performance. 3) The good applicability of the proposed method. We do not need to manually tune the parameters; instead, we can just directly apply the model to the registration task to obtain the reported results, which is flexible in clinical application.

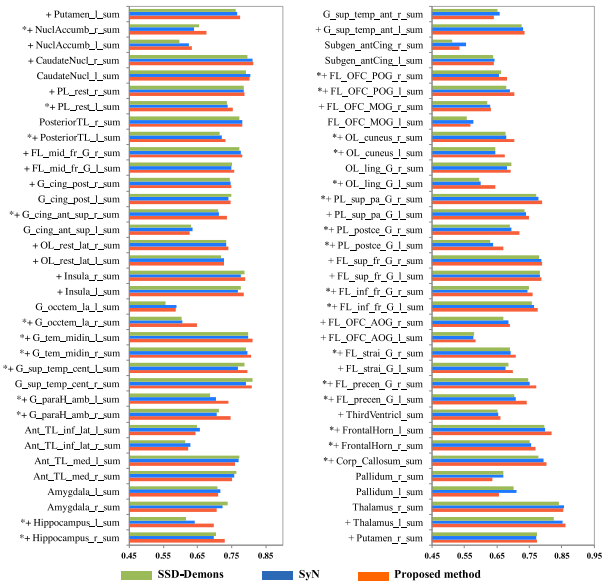

Fig. 10. Comparison of the registration results by **SSD-Demons**, **SyN** and the **proposed method**, repectively. The results are evaluated in term of DSC across the 70 ROIs in IXI database. "+" indicates that the proposed method outperforms the two state-of-the art methods and "∗" means statistically significant improvement ($p < 0.05$ for paired t-test).

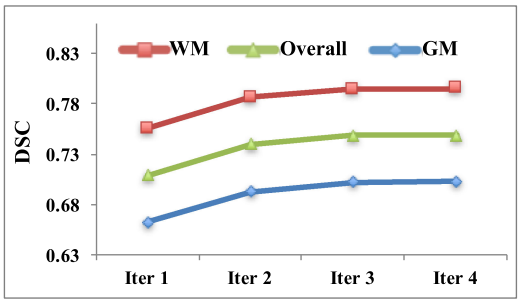

Fig. 11. The registration performance (in DSC) by iterating the trained model for different times.

C. Experiments on ADNI Database

In this section, in order to further increase the challenge of the registration task, we perform the registration *across two different databases*, i.e., LONI and ADNI. As we know, LONI data are the brain images scanned from young adults, while ADNI data are from the old adults containing Alzheimer's disease subjects. In this case, the reference-subject image pair may have very large appearance and anatomical variation, as shown in the first and last columns in Fig. 16. For the LONI data (excluding the training images), we randomly select 4 images to serve as the reference image. All 50 images in ADNI are registered to the 4 references. So, we totally perform 200 times registration for each method. In this section, we first evaluate each contribution of this paper, and then compare with the state-of-the-art methods.

Although we saturate the displacement magnitude for feasible training, we can still estimate the large deformations by applying the trained model iteratively in the application stage. The displacement magnitude can thus be iteratively and accurately accumulated to the actual large deformation. As shown in Fig. 11, the registration performance has improved significantly after applying the model for the second iteration. The performance reaches convergence after the third iteration, as the incremental deformations are mostly vanished.

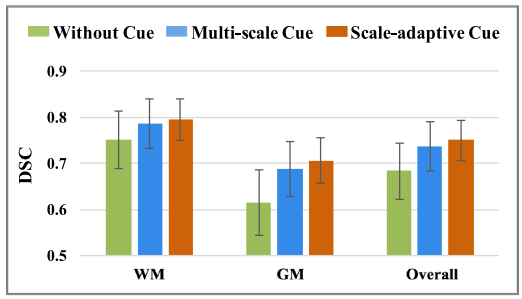

Fig. 12. Mean DSC value with standard deviation evaluated on the WM and GM tissue maps, after performing the deformable registration by using the models trained without the contextual cue (**Without Cue**), with multi-scale contextual cue (**Multi-Scale Cue**), and using the proposed scale-adaptive contextual cue (**Scale-adaptive Cue**).

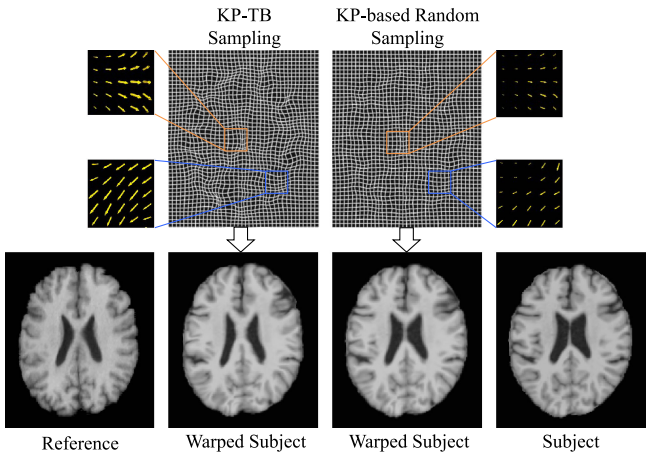

Fig. 13. Visual comparison of registration results using the proposed cue-aware deep network under different sampling strategies in the **training stage**: the KP-based random sampling and the proposed KP-TB sampling.

1) Evaluation for the Contribution of the Auxiliary Contextual Cue: We evaluate the contribution of the auxiliary contextual cue, i.e., the local similarity map, when constructing the registration network. As shown in Fig. 12, without using the contextual cue, the registration performance drops in average along with the increased standard deviation. By using the multi-scale contextual cue, the registration performance is improved. The best performance is achieved using our proposed scale-adaptive contextual cue generated by pooling^C. The results can well demonstrate that, the contextual cue can enhance the awareness of the network for the complex registration task. Moreover, the scale-adaptive cue generated by pooling^C can help suppress the redundant information compared to the case of directly using multi-scale similarity maps directly, thus further improving the registration performance. The results above are also consistent with the loss curves shown in Fig. 7.

2) Evaluation for the Contribution of the Sampling Strategy: Fig. 13 compares the deformation fields using the two models trained under the proposed KP-TB sampling and the KP-based random sampling strategies, respectively. Specifically, in

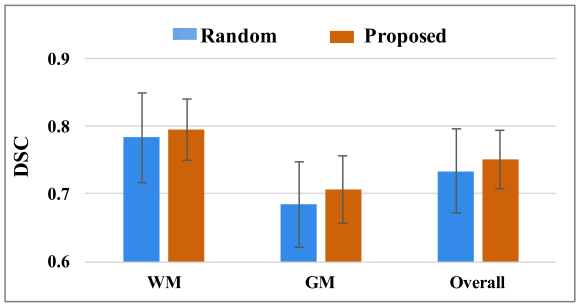

Fig. 14. Mean DSC value with standard deviation evaluated on the WM and GM tissue maps after performing the deformable registration using the same trained model, while with different sampling strategies in the **application stage**: random sampling (**Random**) and the proposed key-points sampling (**Proposed**).

the KP-based random sampling, we acquire patch samples from edges yet disable balanced sampling. The application stage is the same for these two models. As we can observe in the figure, without using the proposed balanced sampling strategy, the deformation field is underestimated. That is, the KP-TB sampling strategy is effective to enhance the accuracy and the generalization capability of the registration network, since the trained model is adaptive to large displacement magnitudes.

To evaluate the influence of the sampling strategy in the application stage, we here compare the random sampling (**Random**) with the proposed KP sampling (**Proposed**) using the same trained model, as the results shown in Fig. 14, the key points sampling is more effective than the random sampling. Since the key points are often located at anatomical rich region like strong edges or corners, thus can generate more distinctive similarity map and can also largely mitigate the ambiguous matching. Moreover, the key points also propagate the accurate displacement estimation to the neighboring smooth region during interpolation, which can eventually contribute to the accurate deformation field.

3) Comparison With the State-of-the-Art Methods: In this section, we use LCC-Demons as the comparison method since it is accurate and robust when registering the images with large appearance/intensity variance. Since we only have WM and GM tissue segmentation maps, we evaluate the performances on the two tissue maps and report the results in Fig. 15. To further illustrate the effectiveness of the proposed method, we also provide visual inspection in Fig. 16.

From Fig. 15, we can observe that, by directly applying the trained registration model in this challenging task, the proposed method achieves the best performance in terms of both DSC and ASD values. This suggests that the proposed method can consistently work well for this challenging registration task. Without any parameter tuning, the proposed method significantly outperforms the two state-of-the-art registration methods by directly applying the trained model to the registration task.

We further provide visual comparison results in both cross-sectional and 3D rendering view, in order to show the detailed differences among these three methods in Fig. 16. From the cross-sectional views in the top two rows, it is obvious that the ventricle regions are more accurately registered to the reference images by adopting the proposed method, as indicated

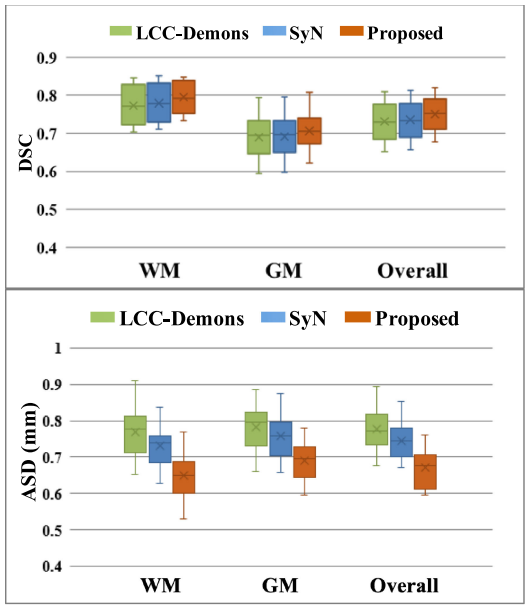

Fig. 15. DSC and ASD (mm) evaluated on the WM and GM tissue maps after performing the deformable registration by **LCC-Demons**, **SyN** and the **Proposed** method, resectively.

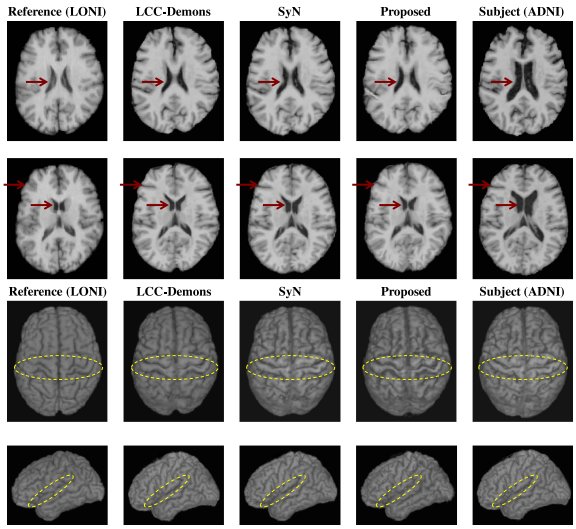

Fig. 16. Qualitative comparison of results by **LCC-Demons**, **SyN** and the proposed method (**Proposed**) in both cross-sectional view (the top two rows) and 3D rendering view (the bottom two rows).

by the red arrows in the figure. Furthermore, more impressive improvements can be obtained on brain cortical regions in the 3D rendering results. For example, in the third row, after registration by the proposed method, the structure of the post central gyrus and the pathway of the central sulcus are more similar to the corresponding reference cortical regions. In the fourth row, the improvements on the lateral fissure (located between the frontal lobe and temporal lobe) are also visibly clear by the proposed method, compared with both LCC-Demons and SyN.

From these results, we can draw the following conclusions. (1) The key-points sampling strategy plays a positive role in better registration of the brain cortical region. Based on our proposed strategy, the key points are more likely to be located at the roots of the sulcus, the crowns of gyrus, or the strong

boundaries. These locations are reliable and important to steer the accurate deformable registration, since they are always of great anatomical significance. (2) The proposed method is accurate and robust even dealing with the challenging registration task in this section, which suggests good generalizability of the trained model based on the proposed cue-aware deep regression network. (3) The proposed method is flexible for clinical application, since it can be consistently performed well for various registration tasks without parameter tuning and setting.

IV. Discussion and Conclusion

We first discuss the runtime of our method. Our algorithm is implemented on an Nvidia Titian XP GPU for both the training and the application stages. In the training stage, usually $6 \sim 8$ epochs are needed for convergence, as shown by the loss curve in Fig. 7(c). We stop the training process when the validation loss cannot decrease. Thus, the training often takes $24 \sim 36$ hours in total. In the application stage, we divide the entire image into 8 non-overlapping blocks. The displacements of the key points are predicted simultaneously for all blocks. Next, we use block-wise TPS (also 8 blocks but with overlap) to interpolate the deformation field for the whole image. Therefore, registering a pair of images often takes $5 \sim 6$ minutes by iterating the trained model for two times (and $7 \sim 8$ minutes for three times) in the application stage. In our future work, we will try to train an interpolation model based on fully convolutional neural network (FCN), which may further speed up the runtime of our registration algorithm.

The auxiliary contextual cue, *i.e.*, the local similarity map, has played an important role to establish an accurate and robust registration model by deep learning. Basically, the local similarity map is a kind of intrinsic hint for conducting the local matching in deformable registration. We calculate this local similarity map through the proposed data-driven convolution and Pooling^{C} operation. In the data-driven convolution, for each patch pair, the kernel (i.e., the small region as shown in Fig. 4) is extracted from the patch sample with different scales. Obviously, the appearance of the kernel varies based on the patch samples. Thus, it is difficult to directly learn the features based on the common kernels as in the conventional convolution operation in CNN. Moreover, this local similarity map can improve the robustness and generalization of the network, which can better fulfill the data diversity. As the images may have inconsistent appearance across different databases, the local similarity map can provide a robust guidance to make the network to be well aware of the local matching during the training of the deformable registration network.

The KP-TB sampling strategy is also another important strategy in this paper. *First*, the key-points (KP) sampling has been proposed for addressing the ambiguous matching problem, which has been illustrated in Section II-B and Fig. 2. Note that, the smooth region without sufficient anatomical details cannot accurately establish the local matching, especially for the patch-wise training manner. While the key points sampling strategy used in both training and testing stages can guarantee that all sampled patches have sufficient anatomical details, which can provide more accurate local matching results. *Second*, the truncated-balanced (TB) sampling has been proposed and only applied in the training stage. This is because, the displacement distribution is quite unbalanced for the real deformation field, as shown in Fig. 2. The TB sampling can make the network adaptive to different displacement magnitudes, which can improve the network accuracy during training, and eventually predict accurate displacement vector in the testing stage.

TABLE II
The Registration Results After Performing the Proposed Method **(Proposed)** and After Each Step Used to Generate the Training Data (**Aft. Step 1**, **Aft. Step 2**, **Aft. Step 3**)

	DSC (%)		
	GM	WM	Overall
Aft. Step 1	69.36±5.39	78.16±6.16	73.76±5.11
Aft. Step 2	69.88±5.21	79.23±5.11	74.56±4.92
Aft. Step 3	**71.02**±5.07	**80.54**±4.52	**75.78**±4.30
Proposed	70.49±4.93	80.02 4.46	75.26±4.37

The ground-truth deformations used to train the registration model have been carefully prepared based on existing registration algorithms, and also with the help of accurate tissue segmentation. Here, we discuss the registration performance with regard to the quality of the training data. Specifically, we have randomly selected 10 pairs of images (the image pairs used in Section III-C) that are not included in our previous training. Then, for each pair, we process it through the three steps that we used to generate the training data, which have been illustrated in Section III. Then, these registration results are compared with the results obtained by our proposed method. The DSCs after performing three respective steps are reported in the Table II, in addition to the results using our proposed method to directly register the image pairs. From the results we can observe that, the performance of our method is restricted by the upper bound in preparing the training data. However, Step 3 actually considers manual tissue segmentation, while our method outperforms Step 2 where only the intensity image is available. That is, our method performs better than the conventional registration method (even after manual yet tedious parameter tuning) in the application stage where only intensity information can be used for guiding the registration.

In this paper, we have proposed a novel deformable registration method of using the deep neural network to directly learn the mapping from an image pair to the corresponding deformation field. This highly non-linear and complex mapping was modeled by the novel cue-aware deep regression network, in which we adopted contextual cue to better guide the learning process. Due to ambiguous matching and unbalanced deformation distribution, a key-point truncated-balanced sampling strategy was developed to generate an informative and well-distributed training set to facilitate learning. Experiments on variable databases and registration tasks have shown improved accuracy and robustness, which could be applicable to various clinical applications in the future.

References

[1] D. Shen *et al.*, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, 2017.
[2] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[4] H.-I. Suk *et al.*, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[5] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[6] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.

[7] J. M. Wolterink *et al.*, "Deep MR to CT synthesis using unpaired data," in *Proc. Int. Workshop Simul. Synth. Med. Imag.*, Sep. 2017, pp. 14–23.

[8] M. Kim *et al.*, "A general fast registration framework by learning deformation—Appearance correlation," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1823–1833, Apr. 2012.

[9] Q. Wang *et al.*, "Predict brain MR image registration via sparse learning of appearance and transformation," *Med. Image Anal.*, vol. 20, no. 1, pp. 61–75, 2015.

[10] X. Yang *et al.*, "Quicksilver: Fast predictive image registration–A deep learning approach," *J. NeuroImage*, vol. 158, pp. 378–396, 2017.

[11] H. Sokooti *et al.*, "Nonrigid image registration using multiscale 3-D convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2017, pp. 232–239.

[12] J. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, 1998.

[13] A. Sotiras *et al.*, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.

[14] H. Lester and S. R. Arridge, "A survey of hierarchical nonlinear medical image registration," *Pattern Recognit.*, vol. 32, no. 1, pp. 129–149, 1999.

[15] D. L. Hill *et al.*, "Medical image registration," *Phys. Med. Biol.*, vol. 46, no. 3, pp. R1–R45, 2001.

[16] D. Rueckert and J. A. Schnabel, "Medical image registration," in *Biomedical Image Processing*. New York, NY, USA: Springer, 2010, pp. 131–154.

[17] B. Zitova and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[18] M. Holden, "A review of geometric transformations for nonrigid body registration," *IEEE Trans. Med. Imag.*, vol. 27, no. 1, pp. 111–128, Jan. 2008.

[19] M. A. Viergever *et al.*, "A survey of medical image registration—Under review," *Med. Image Anal.*, vol. 33, pp. 140–144, 2016.

[20] J.-P. Thirion, "Image matching as a diffusion process: An analogy with Maxwell's demons," *Med. Image Anal.*, vol. 2, no. 3, pp. 243–260, 1998.

[21] T. Vercauteren *et al.*, "Diffeomorphic demons: Efficient nonparametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[22] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, S. M. Smith. FSL. *NeuroImage*, vol. 62, pp. 782–90, 2012.

[23] R. P. Woods *et al.*, "Automated image registration: I. General methods and intrasubject, intramodality validation," *J. Comput. Assisted Tomography*, vol. 22, no. 1, pp. 139–152, 1998.

[24] R. P. Woods *et al.*, "Automated image registration: II. Intersubject validation of linear and nonlinear models," *J. Comput. Assisted Tomography*, vol. 22, no. 1, pp. 153–165, 1998.

[25] P. Hellier *et al.*, "Inter-subject registration of functional and anatomical data using SPM," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2002, pp. 590–597.

[26] B. A. Ardekani *et al.*, "Quantitative comparison of algorithms for intersubject registration of 3-D volumetric brain MRI scans," *J. Neurosci. Methods*, vol. 142, no. 1, pp. 67–76, 2005.

[27] D. L. Collins and A. C. Evans, "Animal: Validation and applications of nonlinear registration-based segmentation," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 11, no. 8, pp. 1271–1294, 1997.

[28] D. Shen and C. Davatzikos, HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 11, pp. 1421–1439, Nov. 2002.

[29] G. Wu *et al.*, "S-HAMMER: Hierarchical attribute-guided, symmetric diffeomorphic registration for MR brain images," *Human Brain Mapping*, vol. 35, no. 3, pp. 1044–1060, 2014.

[30] J. P. Pluim *et al.*, "Image registration by maximization of combined mutual information and gradient information," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2000, pp. 452–461.

[31] J. P. Pluim *et al.*, "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 *et al.*, "Multimodal volume registration by maximization of mutual information," *Med. Image Anal.*, vol. 1, no. 1, pp. 35–51, 1996.

[33] F. Maes *et al.*, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2 pp. 187–198, Apr. 1997.

[34] B. B. Avants *et al.*, "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, 2008.

[35] D. Rueckert *et al.*, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[36] Y. Ou *et al.*, "DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting," *Med. Image Anal.*, vol. 15, no. 4, pp. 622–639, 2011.

[37] M. F. Beg *et al.*, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 139–157, 2005.

[38] G. Wu *et al.*, "TPS-HAMMER: Improving HAMMER registration algorithm by soft correspondence matching and thin-plate splines based deformation interpolation," *NeuroImage*, vol. 49, no. 3, pp. 2225–2233, 2010.

[39] M. Lorenzi *et al.*, "LCC-Demons: A robust and accurate symmetric diffeomorphic registration algorithm," *NeuroImage*, vol. 81, pp. 470–483, 2013.

[40] P. Hellier *et al.* "Intersubject registration of functional and anatomical data using SPM," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2002, pp. 590–597.

[41] B. Glocker *et al.*, "Dense image registration through MRFs and efficient linear programming," *Med. Image Anal.*, vol. 12, no. 6, pp. 731–741, 2008.

[42] Y. Ou *et al.*, "Comparative evaluation of registration algorithms in different brain databases with varying difficulty: Results and insights," *IEEE Trans. Med. Imag.*, vol. 33, no. 10, pp. 2039–2065, Oct. 2014.

[43] A. Klein *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.

[44] L. Wei *et al.*, "Learning appearance and shape evolution for infant image registration in the first year of life," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2016, pp. 36–44.

[45] X. Cao *et al.*, "Dual-Core steered nonrigid registration for multimodal images via bi-directional image synthesis," *Med. Image Anal.*, vol. 41, pp. 18–31, 2017.

[46] X. Cao *et al.*, "Learning-based multimodal image registration for prostate cancer radiation therapy," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2016, pp. 1–9.

[47] B. Gutiérrez-Becker *et al.*, "Learning optimization updates for multimodal registration," in *Proc. Med. Image Comput. Comput.-Assisted Intervention*, 2016, no. 3, pp. 19–27.

[48] S. Miao *et al.*, "A CNN regression approach for real-time 2-D/3-D registration," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1352–1363, May 2016.

[49] K. Ma *et al.*, "Multimodal image registration with deep context reinforcement learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2017, pp. 240–248.

[50] G. Wu *et al.*, "Unsupervised deep feature learning for deformable registration of MR brain images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2013, pp. 649–656.

[51] M.-M. Rohé *et al.*, "SVF-Net: Learning deformable image registration using shape matching," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2017, pp. 266–274.

[52] J. Krebs *et al.*, "Robust nonrigid registration through agent-based action learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2017, pp. 344–352.

[53] B. D. de Vos *et al.*, "End-to-end unsupervised deformable image registration with a convolutional neural network," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, Cham, 2017, pp. 204–212.

[54] H. Li and Y. Fan, "Nonrigid image registration using fully convolutional networks with deep self-supervision," unpublished paper, 2017. [Online]. Available: https://arxiv.org/abs/1709.00799

[55] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Med. Image Anal.*, vol. 5, no. 2, pp. 143–156, 2001.

[56] Y. Zhang *et al.*, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.