Scalable High Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning

Guorong Wu, *Member, IEEE*, Minjeong Kim, Qian Wang, Brent C. Munsell, Dinggang Shen†, *Senior Member, IEEE, and for the Alzheimer's Disease Neuroimaging Initiative**

***Abstract*—Feature selection is a critical step in deformable image registration. In particular, selecting the most discriminative features that accurately and concisely describe complex morphological patterns in image patches improves correspondence detection, which in turn improves image registration accuracy. Furthermore, since more and more imaging modalities are being invented to better identify morphological changes in medical imaging data,, the development of deformable image registration method that scales well to new image modalities or new image applications with little to no human intervention would have a significant impact on the medical image analysis community. To address these concerns, a learning-based image registration framework is proposed that uses deep learning to discover compact and highly discriminative features upon observed imaging data. Specifically, the proposed feature selection method uses a convolutional stacked auto-encoder to identify intrinsic deep feature representations in image patches. Since deep learning is an unsupervised learning method, no ground truth label knowledge is required. This makes the proposed feature selection method more flexible to new imaging modalities since feature representations can be directly learned from the observed imaging data in a very short amount of time. Using the LONI and ADNI imaging datasets, image registration performance was compared to two existing state-of-the-art deformable image registration methods that use handcrafted features. To demonstrate the scalability of the proposed image registration framework, image registration experiments were conducted on 7.0-tesla brain MR images. In all experiments, the results showed the new image registration framework consistently demonstrated more accurate registration results when compared to state-of-the-art.**

***Index Terms*—Deformable image registration, deep learning, hierarchical feature representation.**

Submitted to IEEE Transaction on Biomedical Engineering on March 14, 2015. † Corresponding author.

G. Wu, M. Kim, and D. Shen are with the Department of Radiology and BRIC, the University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 USA. D. Shen is also with Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea. (e-mail: {grwu, mjkim, dgshen}@med.unc.edu).

Q. Wang is with the Med-X Research Institute of Shanghai Jiao Tong University, Shanghai, 200237, China. (e-mail: wangqian@sjtu.edu.cn).

B. Munsell is with the Department of Computer Science, the College of Charleston, Charleston, SC 29424, USA. (e-mail: munsellb@cofc.edu)

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

I. Introduction

DEFORMABLE image registration is very important to neuroscience and clinical studies for normalizing individual subjects to the reference space [1-5]. In deformable image registration, it is critical to establish accurate anatomical correspondences between two medical images. Typically, a patch-based correspondence detection approach is often used, where a patch is a fixed-size symmetric neighborhood of pixel intensity values centered a point in the image. And if two different patches, from two different images, show similar morphological patterns, the two points (at each patch center) are considered to be well corresponded. Therefore, to improve correspondence detection, the problem becomes the one related to feature selection, i.e., how to *consistently* select a set of highly discriminative features that can *accurately*, and *concisely*, capture the morphological pattern presented in the image patch.

Intensity-based feature selection methods are widely used in medical image registration [6-11], however, two image patches that show similar, or even the same, distribution of intensity values do not guarantee the two points are corresponded from an anatomical point of view [4, 12-14]. Handcrafted features, such as geometric moment invariants [15] or Gabor filters [16], are also widely used by many state-of-the-art image registration methods [4, 11, 13, 14, 17]. In general, the major pitfall of using handcrafted features is that the developed model tends to be very ad-hoc. That is, the model is only intended to recognize image patches specific to an image modality or a certain imaging application [18].

Supervised learning-based methods have been proposed to select the best set of features from a large feature pool that may include plenty of redundant handcrafted features [18-24]. However, for this approach, the ground-truth data with known correspondences across the set of training images is required. Because human experts are typically needed to generate ground-truth data, it is well known that obtaining this type of data can be a very laborious and subjective process. In many cases, ground-truth data is simply not available, and even if it does exist, the size of the training population may be very

small, which may dramatically affect the accuracy of the registration method. In general, image registration methods that use supervised learning for feature selection [18, 19] are biased by the registration uncertainty due to the lack of ground truth.

Because deformable image registration is very specific to the input data, it typically takes months, or even years, to develop a new image registration method that has acceptable performance for a new imaging modality or new imaging application. The conventional way of selecting features, including the development of a similarity measurement, requires expert knowledge that is directly related to modality and application. For instance, it is not straightforward to apply the same feature selection methods specifically designed for 1.5-tesla T1 weighted MR image to 7.0-tesla T1 weighted MR images due to the significantly high signal-to-noise ratio in the 7.0-tesla data [25]. Meanwhile, handcrafted features are expensive because manually intensive efforts are required to tune the model for a particular medical image registration application. With the rapid progression of imaging technologies, more and more new modalities are emerging with potentials in disease diagnosis and treatment. Thus the need for a general image registration framework that can quickly be deployed to new modalities and new applications is highly desirable.

To overcome the limitations mentioned above, unsupervised learning-based feature selection methods require further investigation. Because of the complexity of the data, conventional unsupervised approaches that use simple linear models, such as PCA and ICA [26, 27], are typically not suitable because they are unable to preserve the highly non-linear relationships when projected to the low-dimensional space. More advanced methods, such as ISOMAP [28], kernel SVM [29, 30], locally linear embedding (LLE) [31, 32], and sparse coding [33], can learn the non-linear embedding directly from the observed data within a single layer of projection. However, since the learned features are found using only a single layer, or a *shallow model* [34], the selected features may lack high-level perception knowledge (e.g., shape and context information), and may not be suitable for correspondence detection.

Recently, unsupervised deep learning feature selection techniques have been successfully applied to solve many difficult computer vision problems [30, 34-42]. The general concept behind deep learning is to learn hierarchical feature representations by inferring simple representations first and then progressively build up more complex ones from the previous level. Compared with the shallow models, a deep learning architecture can encode multi-level information from simple to complex. Thus, for image registration, deep learning is very promising because it: **(1)** is an unsupervised learning approach that does not require ground truth, **(2)** uses a hierarchical deep architecture to infer complex non-linear relationships, **(3)** is completely data-driven and not based on handcrafted feature selection, and **(4)** can quickly and efficiently compute the hierarchical feature representation for any image patch in the testing data given the trained hierarchical deep architecture (or network).

In this paper, we propose to learn the hierarchical feature representations directly from the observed medical images by using unsupervised deep learning paradigm. Specifically, we introduce a stacked auto-encoder (SAE) [34, 37, 38, 42] with convolutional network architecture [41, 43] into our unsupervised learning framework. The inputs to train the convolutional SAE are the 3D image patches. Generally speaking, our learning-based framework consists of two components, i.e., the encoder and decoder networks. On one hand, the multi-layer encoder network is used to transfer the high-dimension 3D image patches into the low-dimension feature representations, where a single auto-encoder is the building block to learn non-linear and high-order correlations between two feature representation layers. On the other hand, the decoder network is used to recover 3D image patches from the learned low-dimensional feature representations, acting as feedback to refine the inferences in the encoder network. Since the size of 3D image patches can be as large as $\sim 10^4$, it is very computational intensive to directly use a SAE to learn useful features in each layer. To overcome this problem, we use a convolutional network [41] to efficiently learn the translational invariant feature representations [41, 44] such that the learned features are shared among all image points in a certain region. Finally, we present a general framework to fast develop high performance image registration method by allowing the learned feature representations to steer the correspondence detection between two images.

The main contributions of this paper are two-fold: *First*, in order to accurately recognize complex morphological patterns in 3D medical image patches, a deep learning feature selection method is proposed. The benefits of using deep learning for feature selection are: **(1)** does not require manually labeled ground-truth data (that typically is a laborious, subjective, and error-prone process) so it does not suffer from the same limitations as those found in the supervised methods, and **(2)** offers a hierarchical learning paradigm to learn not only low-level but also high-level features that are more flexible than conventional handcrafted features, or even the best features found by the existing supervised learning-based feature selection methods. *Second*, since all existing state-of-the-art image registration frameworks use supervised learning or handcrafted feature selection methods, they generally do not scale well to the new data. However, because the proposed deep learning feature selection method does not suffer from these limitations, the proposed image registration framework can be quickly deployed to perform deformable image registration on new image modalities or new imaging applications with little to even no human intervention. This work is the extension of our previous work in [45], where we further refine the registration performance by using more advanced convolutional SAE, comprehensively evaluate the registration results w.r.t. current state-of-the-art registration methods, and show the potential of our learning-based registration framework in rapid development of new image registration method on brand new 7.0-tesla MR images.

To assess the performance of the proposed registration

framework, we evaluate its registration performance on the conventional 1.5-tesla MR brain images (i.e., the elderly brains from ADNI dataset and the young brains from LONI dataset [46]) that show the proposed registration framework achieves much better performance than the existing state-of-the-art registration methods with handcrafted features. We also demonstrate the scalability of the proposed registration framework on 7.0-tesla MR brain images, where we rapidly develop an accurate image registration method for this new modality with satisfactory registration results.

The remaining sections are organized as follows: In Section 2 we first present the deep learning approach for extracting intrinsic feature representations and our new learning-based registration framework. In Section 3 we evaluate the performance of the proposed registration framework, and in Section 4 we provide a brief conclusion

II. METHOD

Because the input feature space defined by a 3D image patch is typically very large, likely to contain redundant and spurious information, and have small translation differences, the proposed method addresses these issues by implementing a convolutional stacked autoencoder (SAE) network. In general, the goal of the convolutional SAE to reduce this high dimension feature space to some lower dimension representation defined by a set of intrinsic basis functions that are robust to redundant and spurious data artifacts and invariant to translation differences. Using the learned convolutional SAE each input high dimension feature vector is efficiently transformed into a small set of coefficients (i.e. intrinsic feature representation) that well describes the morphological pattern presented in the 3D image patch

A. Naive Methods for Intrinsic Feature Representations

K-means [47] and Gaussian mixture model (GMM) [48] are the two well-known clustering methods that are based upon linear learning models. In particular, given a set of training data $X_{L\times M}$, where L is the dimension of the data and M is the number of samples, the clustering methods learn K centroids such that each sample can be assigned to the closest centroid. Suppose the observed feature vectors (e.g. image intensity patches) form a feature space and the appropriate K centroids in the high-dimension feature space are known. A typical pipeline is to define a function $f: \mathcal{R}^L \to \mathcal{R}^K$ that map the observed L-dimension feature vector to a K-dimension feature vector ($K < L$) [49]. For instance, we could first calculate the affiliations for each observed feature vector (w.r.t. to the K centroids) and then use such affiliations as morphological signatures to represent each key point in the feature space. However, the limitation of K-means and GMM is that the number of centroids is required to be very large as the input dimension grows [42]. Thus, these clustering methods may not be applicable in learning the intrinsic representations for high-dimension medical images.

PCA [50] is one of the most commonly used methods for dimension reduction. PCA extracts a set of basis vectors from the observed data, which maximize the data variance of the projected subspace (spanned by the basis vectors). The basis vectors are obtained by calculating the eigenvectors of the covariance matrix of the input data. Given the observed data $X = [x_1, \dots, x_m, \dots, x_M]$, the following steps are sequentially applied in the training stage: (1) calculate the mean by $\hat{x} = \frac{1}{M}\sum_{m=1}^{M} x_m$; (2) compute the eigenvectors $E = \left[e_j\right]_{j=1,\dots L}$ for the covariance matrix $\frac{1}{M-1}\bar{X}\bar{X}^T$, where $\bar{X} = [x_m - \hat{x}]_{m=1,\dots,M}$ and E are sorted in the decreasing order of eigenvalues; (3) determine the first Q largest eigenvalues such that $\sum_{j=1}^{Q}(\lambda_j)^2 > f_Q \sum_{j=1}^{L}(\lambda_j)^2$, where f_Q defines the proportion of the remaining energy. In the end, each training data x_m can be approximately reconstructed as $x_m = \hat{x} + E_Q b$, where E_Q contains the first Q largest eigenvectors of E and $b = E_Q^T(x - \hat{x})$. In the testing stage, give the new testing data x_{new}, its low-dimensional feature representation can be obtained by $b_{new} = E_Q^T(x_{new} - \hat{x})$. This classic approach for finding low-dimension representations has achieved many successes in medical image analysis area [51, 52]. However, PCA is only an orthogonal linear transform and is not optimal for finding structures with highly non-Gaussian distributions. As we show in the experiment section, such feature representations learned by PCA can hardly assist image registration to establish more accurate correspondences in feature matching.

Since there are huge variations of anatomical structures across individual medical images, the above unsupervised learning methods have limitations in finding the intrinsic feature representations [30, 35, 53]. In the following, we present the new unsupervised deep learning paradigm to infer the intrinsic feature representations from a set of observed 3D image patches

B. Learn Intrinsic Feature Representations by Unsupervised Deep Learning

Introduction of Auto-Encoder. Auto-Encoder (AE) is one typical neural network and structurally defined by three sequential layers: the input layer, the hidden layer, and the output layer. Here, the goal of AE is to learn the latent feature representations from the 3D image patches collected from medical images. Let D and L denote, respectively, the dimensions of hidden representations and input patches. Given an input image patch $x_m \in \mathcal{R}^L$ ($m = 1, \dots, M$), AE maps it to be an activation vector $h_m = [h_m(j)]_{j=1,\dots,D}^T \in \mathcal{R}^D$ by $h_m = f(Wx_m + b_1)$, where the weight matrix $W \in \mathcal{R}^{D\times L}$ and the bias vector $b_1 \in \mathcal{R}^D$ are the encoder parameters. Here, f is the logistic sigmoid function $f(a) = 1/\left(1 + \exp(-a)\right)$. It is worth noting that h_m is considered as the representation vector of the particular input training patch x_m via AE. Next, the representation h_m from the hidden layer is decoded to a vector $y_m \in \mathcal{R}^L$, which approximately reconstructs the input image patch vector x by another deterministic mapping $y_m = f(W^T h_m + b_2) \approx x_m$, where the bias vector $b_2 \in \mathcal{R}^L$ is the decoder parameters. Therefore, the energy function in AE can be formulated as:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2496253, IEEE Transactions on Biomedical Engineering

$$\{W, b_1, b_2\} = \arg\min_{W,b_1,b_2} \sum_{m=1}^{M} \left\| f\left(W^T\left(f(Wx_m + b_1)\right)\right) + b_2 - x_m \right\|_2^2. \quad (1)$$

The sparsity constraint upon the hidden nodes in the network usually leads to more interpretable features. Specifically, we regard each hidden node $h_m(j)$ as being "active" if the value of $h_m(j)$ is close to 1 or "inactive" if the degree is close to 0. Thus, the sparsity constraint requires most of the hidden nodes to remain "inactive" for each training patch x_m. Specifically, the Kullback-Leibler divergence [37, 40, 42] is used to impose the sparsity constraint to each hidden node by enforcing the average activation degree over the whole training data, i.e., $\hat{\rho}_j = \frac{1}{M}\sum_{m=1}^{M} h_m(j)$, to be close to a very small value ρ (here, ρ is set to 0.001 in the experiments):

$$KL\left(\rho \| \hat{\rho}_j\right) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho)\log \frac{1-\rho}{1-\hat{\rho}_j}. \quad (2)$$

Then, the overall energy function of AE with sparsity constraint is defined as:

$$\{W, b_1, b_2\} = \arg\min_{W,b_1,b_2} \sum_{m=1}^{M} \left\| f\left(W^T\left(f(Wx_m + b_1)\right)\right) + b_2 - x_m \right\|_2^2 + \beta \sum_{j=1}^{D} KL\left(\rho \| \hat{\rho}_j\right), \quad (3)$$

where β controls the strength of sparsity penalty term. Typical gradient based back-propagation algorithm can be used for training single AE [34, 35].

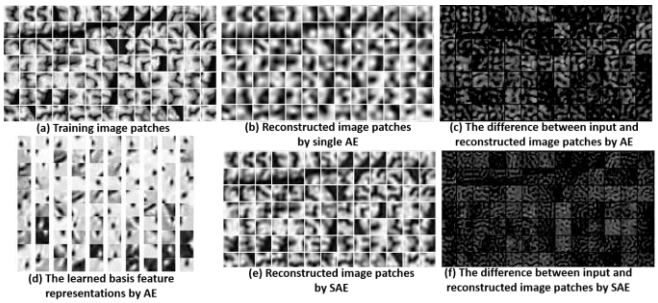


Fig. 1. The reconstructed image patches by single Auto-Encoder (b) and Stacked Auto-Encoder (e). The bright and dark colors indicate large and small reconstruction errors, respectively.

Stacked Auto-Encoder. A single AE is limited in what it can present, since the model is shallow in learning. As shown in Fig. 1(a), a set of training image patches are sampled from brain MR images, each sized at 15×15 (For demonstration, we use 2D patches as examples). We set the number of hidden nodes to be 100 in this single AE. The learned basis feature representations are shown in Fig. 1(d), where a 2D slice represents part of 3D filter. Most of them look like Gabor filters that can detect edges in different orientations. The reconstructed image patches are shown in Fig. 1(b). It is obvious that many details have been lost after the reconstruction from low-dimension representations, as the bright regions displayed in Fig. 1(c).

The power of deep learning emerges when several AEs are stacked to form a Stacked Auto-Encoder (SAE), where each AE becomes a building block in the deep learning model. In order to train SAE, a greedy layer-wise learning approach [36, 37] is used to train a single AE. Specifically, three steps are need to train an SAE, i.e., (1) pre-training, (2) unrolling, and (3) fine-tuning [38]. In the pre-training step, we train the 1st AE with all image patches as the input. Then, we train the 2nd AE by using the activations $h^{(1)}$ of the 1st AE (pink circles in Fig. 2) as the input. In this way, each layer of features captures strong, high-order correlations based on outputs from the layer below. This layer-by-layer learning can be repeated for many times. After pre-training, we build a deep learning network by stacking the AE in each layer, with the higher layer AE nested within the lower layer AE. Fig. 2 shows a SAE consisting of 2-layer stacked AEs. Since the layer-by-layer pre-training procedure provides very good initialization for the multi-level network, we can efficiently use the gradient-based optimization method (such as L-BFGS or Conjugate Gradient [54]) to further refine the SAE parameters in the fine-tuning stage. Due to the deep and hierarchical nature of the network structure, a SAE can discover highly non-linear and complex feature representations for patches in medical images. As shown in Fig. 1(e) and (f), the patch reconstruction performance by SAE becomes much better than using a single AE, where the SAE consists of only 2 layers and the numbers of hidden nodes in each layer are 255 and 100, respectively. It is worth noting that the reconstruction errors in Fig. 1(c) and (f) are in the same range, where bright and dark colors indicate large and small reconstruction errors, respectively.

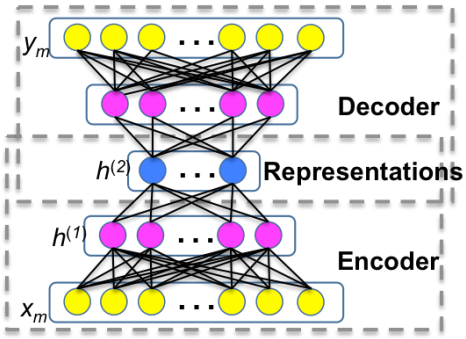


Fig. 2. The hierarchical architecture of Stacked Auto-Encoder (SAE).

Convolutional SAE network. Due to complex nature of medical images, learning the latent feature representations in medical data by employing deep learning is much more difficult than similar applications in computer vision and machine learning areas. In particular, the dimension of input training patch is often very high. For example, the intensity vector of a $21 \times 21 \times 21$ 3D image patch has 9,261 elements. Thus, the training of SAE network becomes very challenging. To alleviate this issue, we resort to using convolutional technique to construct the SAE network.

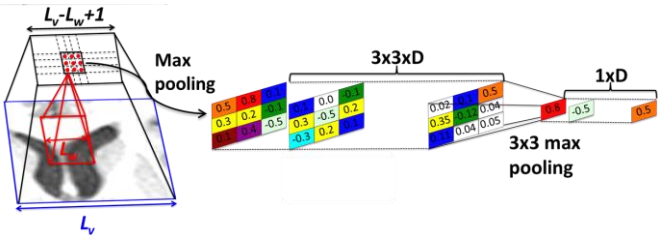


Fig. 3. The 3×3 max pooling procedure in convolutional network.

As shown in Fig. 3, the input to the convolutional SAE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2496253, IEEE Transactions on Biomedical Engineering

network is the large image patch $\mathcal{P}_v$ with patch size L_v. To make it simple, here, we explain the convolutional SAE network with 2D image patch as example. Since the dimension of the image patch $\mathcal{P}_v$ is too large, we let a $L_w \times L_w$ ($L_w < L_v$) sliding window $\mathcal{P}_w$ (red box in Fig. 2) go through the entire big image patch $\mathcal{P}_v$, thus obtaining $(L_v - L_w + 1) \times (L_v - L_w + 1)$ small image patches. Eventually, we use these small image patches $\mathcal{P}_w$ to train the auto-encoder in each layer, instead of the entire big image patch $\mathcal{P}_v$. Given the parameters of network (weight matrix W and bias vector b_1 and b_2), we can compute $(L_v - L_w + 1) \times (L_v - L_w + 1)$ activation vectors, where we use the red dots in Fig. 3 to denote the activation vectors in a 3×3 neighborhood. Then max pooling is used to shrink the representations by a factor of C in each direction (horizontal or vertical). The right part of Fig. 3 demonstrates the 3×3 max pooling procedure ($C = 3$). Specifically, we compute the representative activation vector among these 9 activation vectors in the 3×3 neighborhood by choosing the maximum absolute value for each vector element. Thus, the number of activation vector significantly reduces to $\frac{L_v - L_w + 1}{C} \times \frac{L_v - L_w + 1}{C}$. Since we apply the maximum operation, shrinking the representation with max pooling allows high-level representation to be invariant to small translations of the input image patches and reduce the computational burden. This translation invariant advantage is very useful in establishing anatomical correspondences between two images, as we will demonstrate in our experiments.

Sample the image patches. Typically, one brain MR image, with $1 \times 1 \times 1\text{mm}^3$ spatial resolution, has over 8 million voxels in the brain volume. Obviously, there would be too many image patches to train the deep learning network, not to say that we extract the image patches across a set of training images. Therefore, adaptive sampling strategy is necessary to secure not only using an enough number of image patches but also selecting the most representative image patches to learn the latent feature representations for the entire training set.

To this end, there are two criteria for sampling image patches: 1) In a local view, the selected image patches should locate at distinctive regions in the image, such as sulcal roots and gyral crowns in MR brain images, since they are relatively easy to identify their correspondences; 2) In a global view, the selected image patches should cover the entire brain volume, while the density of sampled points could be low in the uniform regions and high in the context-rich regions. To meet these criteria, we use the importance sampling strategy [11] to hierarchically sample the image patches. Specifically, we smooth and normalize the gradient magnitude values over the whole image domain of each training image. Then, we use the obtained values as the importance degree (or probability) of each voxel to be sampled for deep learning. Note that, although more sophisticated method [55] could be used here for guiding sample selection, we use a simple gradient guided strategy since it is computationally fast. Based on this importance (probability) map, a set of image patches can be sampled via Monte Carlo simulation in each image. Fig. 4(a) shows the non-uniform sampling based on the importance (probability) map in learning the intrinsic feature representations for MR brain images. It can be observed from Fig. 4 that the sampled image patches (with the center point of each sampled patch denoted by the red dot in Fig. 4(b)) are more concentrated at the context-rich (or edge-rich) regions, where the values of importance (or probability) are high.

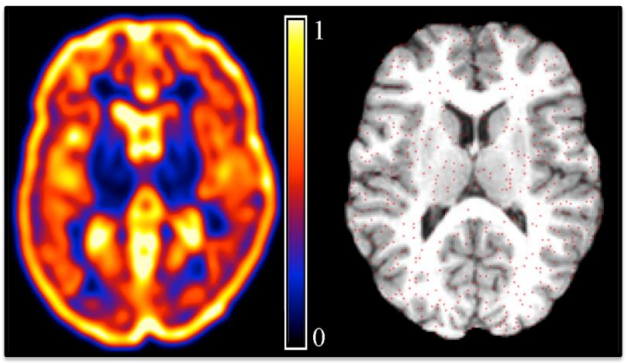


Fig. 4. The importance map and the sampled image patches (denoted by the red dots) for deep learning. The color bar indicates the varying importance values for individual voxels.

C. Learning-based Registration Framework Using Learned Feature Representations

After training the convolutional SAE upon a large amount of 3D image patches, it is efficient to obtain the low-dimensional feature representations (blue circles in Fig. 2) by simple matrix multiplication and addition in each encoder layer. Our implementation is developed based on the deep learning software freely available at University of Toronto (http://deeplearning.cs.toronto.edu/codes). Such low-dimensional feature representation, regarded as the *morphological signature*, allows each point to accurately identify the correspondence during image registration as demonstrated above.

Since the convolutional SAE can directly learn the feature representations from the observed data, the learning procedure is completely free of the limitation of requiring ground-truth data. Thus, it is straightforward to learn optimal feature representations for specific dataset, with little or even no human intervention. Thanks to the state-of-the-art deformation mechanisms developed in many registration methods, we propose a learning-based registration framework by replacing with the learned feature representations and still inheriting the existing deformation mechanism to derive the deformation pathway. In general, deformable image registration methods can be categorized into two types: intensity-based (typical example: Demons [10, 56]) and feature-based (typical example: HAMMER [4, 14, 57]) approaches . In the following, we show how to improve the state-of-the-art registration methods by integrating the feature representations via deep learning.

Multi-Channel Demons With Deep Feature Representation. Many image registration methods, e.g., Demons [10, 56], use the gradient-based optimization

approach to iteratively estimate the deformation fields. To utilize multiple image information, such as multi-modality images, multi-channel Demons was proposed by allowing one channel carrying one information source [58-60]. Therefore, it is straightforward to deploy multi-channel Demons, by regarding all elements of the learned feature representations as the multi-channel information. In each iteration, the update vector field is computed independently for each channel and then averaged to update the deformation field, until the overall feature difference across channels reaches local minima.

HAMMER with Deep Feature Representation. HAMMER registration algorithm (with its ITK-based source code available at http://www.nitrc.org/projects/hammerwml) is a typical feature-based deformable registration for MR brain images. Generally, we can replace the handcrafted attribute vectors (i.e., the low-order statistics in multi-resolution histograms) in the feature-based HAMMER registration method [14, 61] with the learned feature representations by the convolutional SAE. Next, we follow the hierarchical deformation optimization mechanism in HAMMER to estimate the deformation pathway between two images. Specifically, we alternate the following 5 steps until the image registration converges:

1. For each image, we follow the importance sampling strategy explained in Section B to select a small number of key points in both images. Since the key points usually locate the distinct regions in the image, as shown in Fig. 4(b), they can establish the correspondences more reliably than other points. Thus, we allow these key points to steer the estimation of entire deform pathway.
2. For each key point, we identify its anatomical correspondence by matching the learned deep feature representations [4, 57]. Here, we use normalized cross correlation as the similarity measurement between the feature representation vectors of the two different points under comparison.
3. Given the correspondence established on key points, we can interpolate the dense deformation field by using thin-plate splines [62].
4. Relax the selection criterion in Step (1) to allow more key points taking part in the correspondence detection until all image points are used as key points.

III. Experiments

Here we evaluate deformable image registration performance of the proposed image registration framework that uses deep learning for feature selection. For comparison, we set diffeomorphic Demons and HAMMER as the baselines for intensity-based and feature-based registration methods, respectively. Then, we extend the diffeomorphic Demons from a single channel (i.e., image intensity) to multi-channel Demons by adapting the learned feature representations via deep learning to multiple channels, which is denoted by M+DP. Similarly, we modify HAMMER to use the feature representations learned via deep learning, and denote the respective method as H+DP. Since PCA is widely used for unsupervised learning, we apply PCA to infer the latent low-dimensional feature representations for our images. After integrating such low-dimensional feature representations by PCA into multi-channel Demons and HAMMER, we can obtain other two new registration methods, denoted as M+PCA and H+PCA, respectively.

A. Experiments on ADNI Dataset

In this experiment, we randomly select 66 MR images from the ADNI dataset (http://adni.loni.ucla.edu/), where 40 images are used to learn feature representations and another 26 images are used for testing image registration. The preprocessing steps include skull removal [63], bias correction [64], and intensity normalization [65]. For each training image, we sample around 7,000 image patches, where the patch size is set to $21 \times 21 \times 21$. In the following experiment, we follow the practical guide in [66] to train the deep learning network. Specifically, the convolutional SAE consists of 8 layers (stacked with 4 AEs). We only apply the max pooling in the lowest layer with the pooling factor $C = 3$. From the lowest to the highest level, the numbers of hidden nodes in each stacked AE are 512, 512, 256, and 128, respectively. Thus, the dimension is 128 after deep learning algorithm is applied. To keep the similar dimension of learned features by PCA, we set the portion of remaining energy f_Q to 0.7 in this experiment. To avoid overfitting, we divide the whole training samples into mini-batches, each consisting of 100 samples. During training, we monitor the progress of learning and adjust the learning rate by inspecting the reconstruction errors. The sparsity target value $\hat{\rho}_j$ in Eq. 2 is set to 0.01.

Advantages of Feature Representations Learned by Deep Learning Network. The power of feature representations learned by deep learning is demonstrated in Fig. 5. A typical image registration result for the elderly brain images is shown in the top of Fig. 5, where the deformed subject image (Fig. 5 (c)) is far from well registered with the

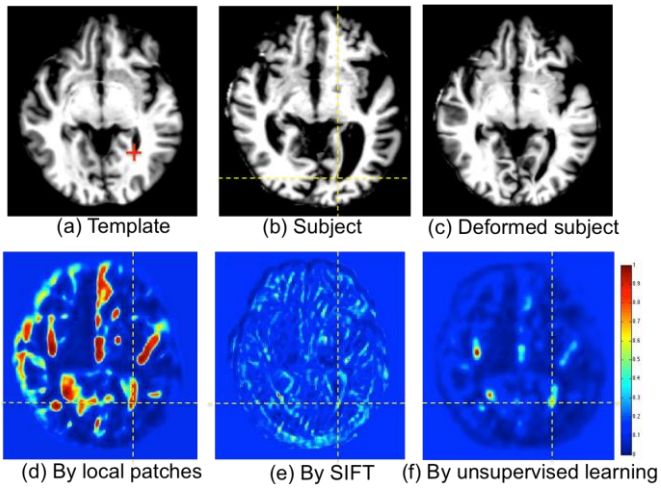


Fig. 5. The similarity maps of identifying the correspondence for the red-crossed point in the template (a) w.r.t. the subject (b) by handcraft features (d-e) and the learned features by unsupervised deep learning (f). The registered subject image is shown in (c). It is clear that the in-accurate registration results might undermine the supervised feature representation learning that highly relies on the correspondences across all training images.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2496253, IEEE Transactions on Biomedical Engineering

template image (Fig. 5(a)), especially for ventricles. Obviously, it is very difficult to learn meaning features given the inaccurate correspondences derived from imperfect image registration, as suffered by many supervised learning methods.

The performance of our learned features is shown in Fig. 5(f). For a template point (indicated by the red cross in Fig. 5(a)), we can successfully find its corresponding point in the subject image, whose ventricle is significantly larger. Each point in Fig. 5(f) indicate its likelihood of being selected as correspondence in the respective location. According to the color bar shown in Fig. 5, it is easy to locate the correspondence of the red-cross template point in the subject image domain, since the high correspondence probabilities are densely distributed right at the corresponding location and then quickly fade away. Other handcrafted features either detect too many non-corresponding points (when using the entire intensity patch as the feature vector as shown in Fig. 5(d)) or have too low responses and thus miss the correspondence (when using SIFT features as shown in Fig. 5(e)). In general, our method reveals the least confusing correspondence information for the subject point under consideration, and implies the best correspondence detection performance that eventually improve the registration accuracy as follows.

Evaluation of Registration Performance. In image registration, one image is selected as the template and the other 25 images are considered as subject images. Before deploying deformable image registration, FLIRT in FSL package (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) is used to linearly align all subject images to the template image. After that, we apply 6 registration methods, i.e., diffeomorphic Demons (simply named as Demons below), M+PCA, M+DP, HAMMER, H+PCA, and H+DP, to normalize those 25 subject images to the template image space, respectively. To quantitatively evaluate the registration accuracy, we first use FAST in FSL package to segment each image into white matter (WM), gray matter (GM), and cerebral-spinal fluid (CSF). After that, we use our in-house tools to label the ventricle (VN) from the CSF segmentation. Here, we use these segmentation results to evaluate the registration accuracy by comparing the Dice ratio of tissue overlap degrees between template and each registered subject image. Specifically, the Dice ratio is defined as:

$$D(R_A, R_B) = \frac{2|R_A \cap R_B|}{|R_A| + |R_B|}, \quad (4)$$

where R_A and R_B denote two ROIs (Regions of Interest) and $|\cdot|$ stands for the volume of the region. The Dice ratios on WM, GM, and VN by 6 registration methods are shown in Table I. It is clear that (1) the registration methods integrated with the feature representations by deep learning consistently outperform the counterpart baseline methods and also the methods using PCA-based feature representations only; (2) H+DP achieves the highest registration accuracy with almost 2.5% improvement in overall Dice ratio against the baseline HAMMER registration method. Since ADNI provides the hippocampus labeling for the template and all 25 subject images, we can further evaluate the Dice overlap ratio on hippocampus. The mean and the standard deviation of the Dice ratios on hippocampus by 6 registration methods are (72.2±3.1)% by Demons, (72.3±2.9)% by M+PCA, (72.5±2.8)% by M+DP, (75.5±2.9)% by HAMMER, (75.6±2.5)% by H+PCA, and (76.8±2.2)% by H+DP, respectively. Compared to the baseline methods, M+DP and H+DP obtain 0.3% and 1.3% improvements in terms of Dice ratios, respectively, where H+DP achieves significant improvement on all WM, GM, CSF tissue overlap ratios under paired *t*-test ($p < 0.05$), as indicated by '*' in Table I. Particularly, the reason of the less improvement by M+DP compared to H+DP might be related with the high number of channels (128 channels) used in M+DP, compared with only less than 10 channels used in [58-60].

TABLE I
THE DICE RATIOS OF WM, GM, AND VN ON ADNI DATASET (UNIT: %)

Method	WM	GM	VN	Overall
Demons	85.7	76.0	90.2	84.0
M+PCA	85.5	76.6	90.2	84.1
M+DP	85.8	76.5	90.9	84.4
HAMMER	85.4	75.5	91.5	84.1
H+PCA	86.5	76.9	91.7	85.0
H+DP	**88.1***	**78.6***	**93.0***	**86.6**

B. Experiments on LONI Dataset

In this experiment, we use the LONI LPBA40 dataset [67] which consists of 40 brain images, each with 56 manually labeled ROIs. We use the first 20 images for learning the latent feature representations and another 20 images for testing the registration performance. The preprocessing procedures include bias correction, intensity normalization and linear registration by FLIRT, which are the same with Section 3.1. For each training image, we sample around 9,000 image patches, where the patch size is again set to $21 \times 21 \times 21$. Other parameters in training convolutional SAE are also the same with Section 3.1. Therefore, the dimension of feature representations after deep learning is 128. To keep the similar dimension of learned features by PCA, we set f_Q to 0.65 in this experiment.

One of the 20 testing images is selected as the template and we apply 6 registration methods to register the rest of 19 testing images to the selected template. The averaged Dice ratio in each ROI by 6 registration methods is shown in Fig. 6. The overall Dice ratios across all 56 ROIs by 6 registration methods are provided in Table 3. Again, H+DP achieves the largest improvement (2.5%) over the baseline HAMMER registration method. Specifically, we perform the paired *t*-test between H+DP and all other 5 registration methods, respectively. The results indicate that H+DP has the statistically significant improvement over all other 5 registration methods in 28 out of 54 ROIs (designated by the red stars in Fig. 6).

C. Experiments on 7.0-Tesla MR Image Dataset

In the previous experiments, we have demonstrated the power of learned feature representations by deep learning in terms of the improved registration accuracy, represented by overlap ratio of structures. As we mentioned early, another

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2496253, IEEE Transactions on Biomedical Engineering

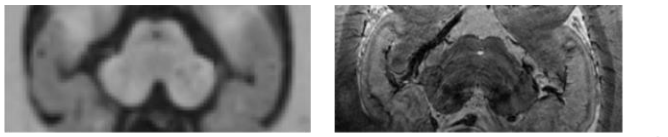
	Demons	M+PCA	M+DP	HAMMER	H+PCA	H+DP
L sup. frontal gyrus	80.2	79.5	79.1	77.3	77.1	78.5
R sup. frontal gyrus	79.7	79.8	80.0	77.5	77.2	78.4
* L middle frontal gyrus	77.4	78.1	78.2	79.5	78.5	82.3
* R middle frontal gyrus	77.0	76.5	77.1	78.2	78.6	80.4
L inf. frontal gyrus	72.2	72.8	72.6	72.8	73.0	74.6
R inf. frontal gyrus	72.0	72.1	72.3	72.6	72.4	74.1
L precentral gyrus	67.9	68.5	68.4	68.6	69.1	71.5
R precentral gyrus	68.6	68.5	69.0	65.0	65.2	65.1
* L middle orbitofrontal gyrus	66.9	66.1	67.1	69.8	69.9	73.5
* R middle orbitofrontal gyrus	66.8	67.5	67.7	69.4	69.5	73.9
* L lateral orbitofrontal gyrus	58.1	58.1	58.5	60.5	61.5	64.9
R lateral orbitofrontal gyrus	55.4	55.6	55.7	64.7	64.1	68.9
L gyrus rectus	66.7	66.1	66.9	67.9	67.2	69.8
R gyrus rectus	68.1	67.7	68.1	65.5	65.0	65.0
L postcentral gyrus	60.5	60.7	61.4	60.5	61.2	63.0
* R postcentral gyrus	62.9	62.4	62.5	63.5	63.1	65.2
* L sup. parietal gyrus	70.7	70.9	70.6	72.6	72.7	74.7
* R sup. parietal gyrus	70.9	70.6	71.2	71.8	71.9	73.5
L supramarginal gyrus	63.8	63.4	64.1	65.1	65.5	67.8
R supramarginal gyrus	63.3	63.7	63.8	65.1	66.4	68.5
* L angular gyrus	63.2	62.8	63.5	66.9	66.8	70.0
R angular gyrus	65.0	65.1	65.7	65.4	65.8	67.5
* L precuneus	65.9	65.7	66.4	70.6	70.9	74.0
* R precuneus	67.3	67.2	67.8	70.8	71.5	76.5
L sup. occipital gyrus	58.1	58.0	58.2	61.2	62.4	65.5
R sup. occipital gyrus	55.4	55.9	56.2	64.5	65.4	67.7
* L middle occipital gyrus	68.7	68.5	68.4	72.6	74.9	80.6
R middle occipital gyrus	67.9	67.4	67.8	71.1	72.0	73.1
L inf. occipital gyrus	67.2	67.8	67.9	65.8	65.5	66.4
R inf. occipital gyrus	66.1	66.5	67.1	62.0	62.0	62.1
L cuneus	63.4	63.4	63.9	64.1	64.5	64.7
* R cuneus	62.2	62.4	62.5	66.0	67.2	70.1
L sup. temporal gyrus	72.5	72.5	72.7	69.5	69.5	70.7
* R sup. temporal gyrus	72.6	73.1	73.4	74.1	74.5	76.4
* L middle temporal gyrus	66.4	66.8	66.8	67.1	66.9	69.5
R middle temporal gyrus	67.9	67.5	67.9	68.3	68.4	69.7
L inf. temporal gyrus	65.6	65.2	65.9	65.8	65.1	66.3
* R inf. temporal gyrus	66.4	66.4	66.5	66.9	67.8	70.0
L parahippocampal gyrus	68.1	68.2	68.5	68.0	69.1	70.1
R parahippocampal gyrus	66.9	66.7	67.2	67.5	69.0	71.0
* L lingual gyrus	69.7	69.8	68.9	69.4	69.2	71.8
* R lingual gyrus	70.6	70.5	70.6	73.6	74.3	77.5
L fusiform gyrus	68.9	68.8	69.1	66.5	66.1	66.2
R fusiform gyrus	68.3	68.3	68.5	67.5	67.5	67.8
L insular cortex	76.4	76.1	76.5	77.5	77.9	79.7
R insular cortex	74.2	74.6	74.7	75.1	76.0	76.1
* L cingulate gyrus	68.1	68.2	68.8	69.5	69.9	71.4
* R cingulate gyrus	67.5	67.4	67.2	69.2	70.5	72.2
* L caudate	73.4	73.4	73.8	74.5	75.0	77.8
* R caudate	73.1	73.0	73.5	76.2	76.4	78.3
* L putamen	76.3	76.5	76.7	77.0	77.7	80.0
* R putamen	76.5	76.3	76.4	76.5	78.6	80.6
* L hippocampus	72.7	72.6	72.8	74.7	75.8	77.7
* R hippocampus	72.8	72.6	73.1	75.9	77.1	81.3
* cerebellum	84.9	85.1	85.9	86.0	87.8	90.3
* brainstem	80.6	81.4	83.1	85.5	86.6	89.1

Fig. 6. The Dice ratios of 56 ROIs on LONI dataset by 6 registration methods.

attractive advantage of deep learning is that we can rapidly learn the intrinsic feature representations for the new imaging modality. In this section, we apply the convolutional SAE to 7.0-tesla MR brain images. The learned feature representations are then integrated with HAMMER, and thus we develop a specific registration method for 7.0-tesla MR brain images.

The advent of 7.0-tesla MR imaging technology [68] enables the achievement of high signal-to-noise ratio (SNR) as well as the dramatically increased tissue contrast compared to the 1.5- or 3.0-tesla MR image. A typical 7.0-tesla MR brain image (with the image spatial resolution $0.35 \times 0.35 \times 0.35mm^3$) is shown in Fig. 7(b), along with a similar slice from a 1.5-tesla scanner (with the resolution of $1 \times 1 \times 1mm^3$) in Fig. 7(a) for comparison. As demonstrated in [69], 7.0-tesla MR image can reveal the brains' architecture with resolution equivalent to that obtained from thin slices *in vitro*. Thus, researchers are able to observe clearly the fine brain structures in μm unit, which was only possible with *in vitro* imaging in the past. Without doubt, 7.0-tesla MR imaging technique has the high potential to be the standard in discovering morphological patterns of human brain in the near future.

Unfortunately, all existing state-of-the-art deformable



(a) 1.5 Tesla MR brain images (1x1x1mm³) (b) 7.0 Tesla MR brain images (0.35x0.35x0.35mm³)

Fig. 7. Large structural difference around hippocampus between 1.5-tesla (a) and 7.0-tesla (b) MR images. The 1.5-tesla image is enlarged w.r.t. the image resolution of the 7.0-tesla image for convenience of visual comparison. .

registration methods, developed for 1.5-tesla or 3.0-tesla MR images, do not work well for the 7.0-tesla MR images, mainly because 1) severe intensity inhomogeneity issue in 7.0-tesla MR images, and 2) much richer texture information than that in 1.5-tesla or 3.0-MR tesla images, as displayed in Fig. 7(b).

Overall 20 7.0-tesla MR images acquired by the method in [68] were used in this experiment, where 10 are used for training by deep learning and another 10 images used for testing the registration performance. We randomly select one image as the template. For the 7.0-tesla scanner (Magnetom, Siemens), an optimized multichannel radiofrequency (RF) coil and a 3D fast low-angle shot (Spoiled FLASH) sequence were utilized, with TR=50ms, TE=25ms, flip angle 10°, pixel band width 30Hz/pixel, field of view (FOV) 200mm, matrix size $512 \times 576 \times 60$, 3/4 partial Fourier, and number of average (NEX) 1. The image resolution of the acquired images is isotropic, e.g., $0.35 \times 0.35 \times 0.35\text{mm}^3$. The hippocampi were manually segmented by neurologist [68]. All images were pre-processed by the following steps: 1) manual skull removal; 2) inhomogeneity correction using N4 bias correction [64]; 3) intensity normalization for making image contrast and luminance consistent across all subjects [65]; 4) affine registration to the selected template by FSL.

For each training image, we sample around 9,000 image patches, where the patch size is set to $27 \times 27 \times 27$. The convolutional SAE consists of 10 layers (stacked with 5 AEs). We only apply the max pooling in the lowest layer with the pooling factor $C = 3$. From low level to high level, the numbers of hidden nodes in each stacked AE are 1024, 512, 512, 256, and 128, respectively. Thus, the dimension of feature representations after deep learning is still 128. In order to achieve the best registration performance, we integrate the learned feature representations trained from 7.0-tesla MR images with the HAMMER registration method.

Several typical registration results on 7.0-tesla MR images are displayed in Fig. 8, where the template and subject images are shown in Fig. 8(a) and (b), respectively. Here, we compare the registration results with diffeomorphic Demons (Fig. 8(c)) and HAMMER (Fig. 8(d)). The registration results by H+DP, i.e., integrating the learned feature representations by deep learning with HAMMER, are display in Fig. 8(e), where the manually labeled hippocampus on template image and the deformed subject's hippocampus by different registration methods are shown by red and blue contours, respectively. Through visual inspection (the overlap of red and blue contours), the registration result by H+DP is much better than both diffeomorphic Demons and HAMMER. Diffeomorphic Demons registration method fails to register 7T image, as shown in Fig. 8 (c), since it is simply driven by image

intensities which suffer from image noise and inhomogeneity in 7T images. In addition, due to huge difference of image characteristics between 7T and 3T images, the hand-crafted features optimized for 3T image also do not work well for 7T images in the feature-based HAMMER registration method either, as shown in Fig. 8(d).

Since we have the manually labeled hippocampus for template and all subject images, we can further quantitatively measure the registration accuracy. The mean and standard deviation of Dice ratios on hippocampus are $(53.5 \pm 4.9)\%$ by diffeomorphic Demons, $(64.9 \pm 3.1)\%$ by HAMMER, and

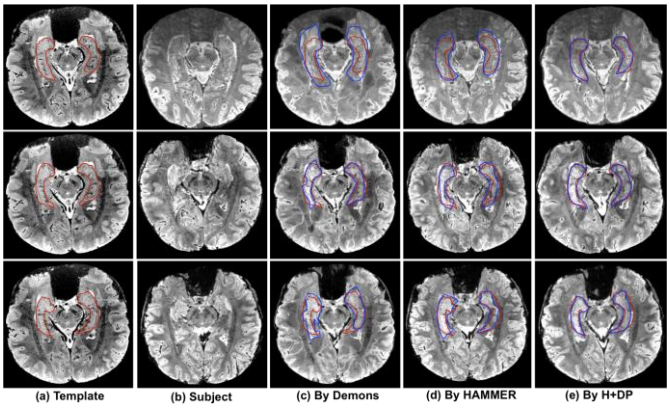


Fig. 8. Typical registration results on 7.0-tesla MR brain images by Demons, HAMMER, and H+DP, respectively. Three rows represent three different slices in the template, subject, and registered subjects.

$(75.3 \pm 1.2)\%$ by H+DP, respectively. Obviously, H+DP achieves a significant improvement on registration accuracy. This experiment demonstrates that (1) the latent feature representations inferred by deep learning can well describe the local image characteristics; (2) we can rapidly develop image registration for the new medical imaging modalities by using deep learning framework to learn the intrinsic feature representations; and (3) the whole learning-based framework is fully adaptive to learn the image data and reusable to various medical imaging applications.

IV. Discussion

A. The Generality of Learned Feature Representations

Recall that we have obtained the feature representations by deep learning on the ADNI dataset in Section III.A. It is interesting to evaluate the generality of deep learning by applying the learned feature representations from the ADNI dataset (which mostly contains elderly brain images) to register the images in the LONI dataset (i.e., containing young brain images). Fig. 9 shows the Dice ratio in each ROI in the LONI dataset by using (1) the baseline HAMMER registration method (in blue), (2) H+DP-LONI (in red) where we use the learned feature representations from LONI dataset and integrate with HAMMER, (3) H+DP-ADNI (in green) where we use the learned feature representations from ADNI dataset and integrate with HAMMER. It is apparent that the registration performance by H+DP-ADNI is comparable to H+DP-LONI, where the average improvements over the baseline HAMMER registration method are 1.99% by H+DP-ADNI and 2.55% by H+DP-LONI, respectively. Under paired *t*-test ($p < 0.05$), we find that H+DP-LONI has 28 ROIs with significant improvement (as indicated by red '*' in Fig. 9), while H+DP-ADNI has 18 ROIs with significant improvement (as indicated by the green '*' in Fig. 9). This indicates that the learned feature representations by the convolutional SAE network are general, although the appearances of two datasets are quite different (i.e., due to aging).

On the other hand, we use the learned feature representations from the LONI dataset to evaluate the registratoin performance on ADNI dataset. Table II shows the tissue overlap ratios by the baseline HAMMER registration method, H+DP-ADNI (learning features from ADNI dataset), and H+DP-LONI (learning features from LONI dataset). Red '*' indicates significat improment over HAMMER under paired *t*-test with $p < 0.05$. It is clear that the feature representations learned from LONI dataset is reusable to ADNI dataset as well, and the registration performance is comparable to the case of directly learning feature reprsentations from the same dataset.

TABLE II
THE DICE RATIOS OF WM, GM, AND VN ON ADNI DATASET (UNIT: %)

Method	WM	GM	VN	Overall
HAMMER	85.4	75.5	91.5	84.1
H+DP-ADNI	88.1*	78.6*	93.0*	86.6
H+DP-LONI	87.7*	78.3	92.8	86.2

B. Computational Time

Table III shows the computational times for all registration methods on ADNI, LONI, and 7.0-tesla datasets. The computation environment is DELL workstation with 8 CPU cores and 16G memory. Multi-thread technique is used in the implementation for each method. Compare to the baseline methods, the additional computational cost mainly comes from the I/O of feature vectors on each image point and also the feature matching procedure, due to high feature dimension.

TABLE III
THE AVERAGE COMPUTATIONAL TIME ON ALL DATASETS (UNIT: MINUTE)

Method	ADNI	LONI	7.0-tesla
Demons	3.0	3.0	16.0
M+PCA	125.0	122.0	260.0
M+DP	125.0	121.0	265.0
HAMMER	15.6	12.6	34.6
H+PCA	106.0	98.0	170.0
H+DP	105.0	98.0	168.0

C. Comparison with Other State-of-the-Art Registration Methods

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2015.2496253, IEEE Transactions on Biomedical Engineering

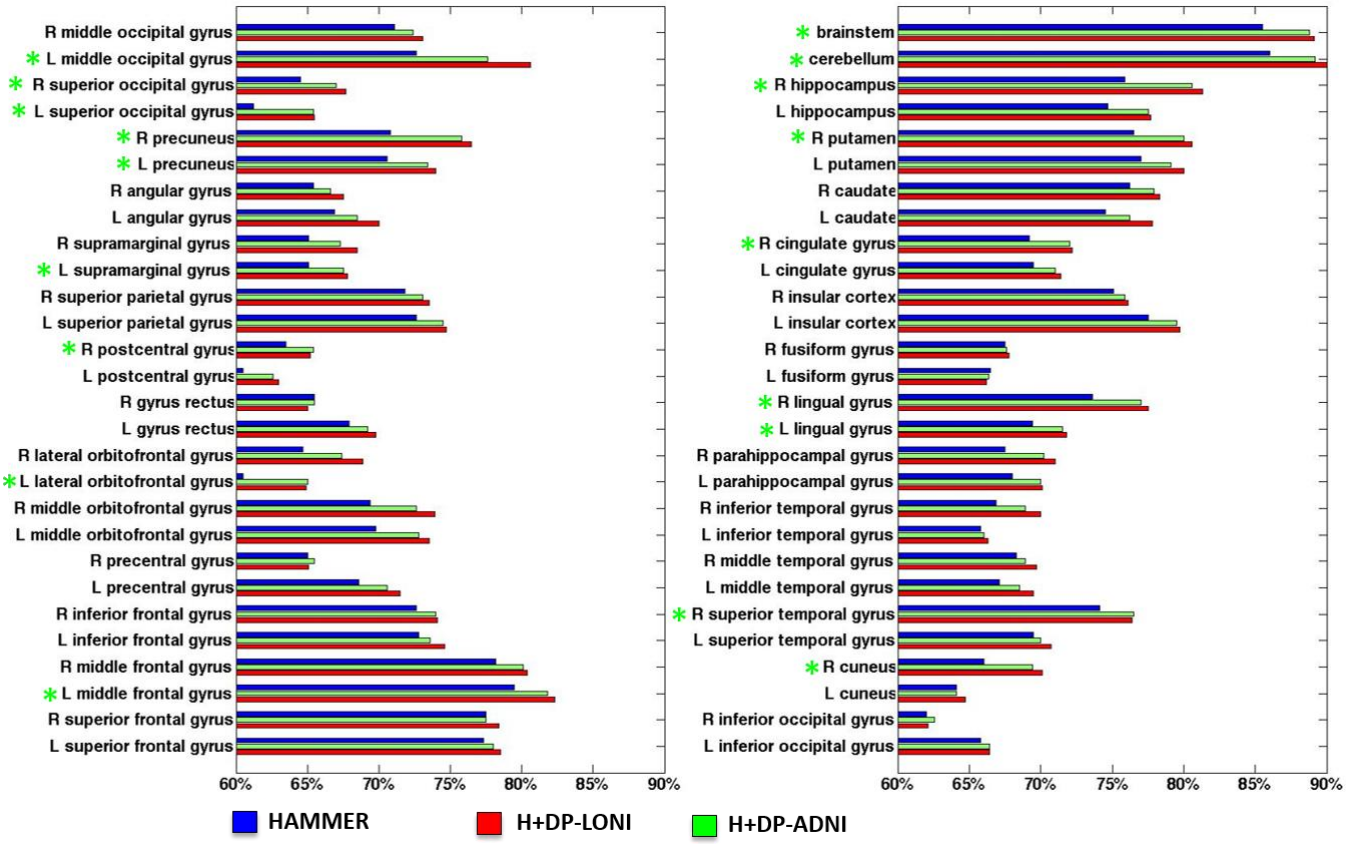


Fig. 9. The Dice ratios of 56 ROIs in LONI dataset by HAMMER (blue), H+DP-LONI (red), and H+DP-ADNI (green), respectively. Note, H+DP-LONI denotes for the HAMMER registration integrating with the feature representations learned directly from LONI dataset, while H+DP-ADNI stands for applying HAMMER registration on LONI dataset but using the feature representations learned from ADNI dataset, respectively.

As we mentioned early, it is not practical to use supervised leaning approaches to find the best feature representations due to the lack of ground truth anatomical correspondences. In our previous work [18, 19], we have to assume the correspondences established by a certain image registration method as the ground truth, which makes the selection of best features not optimal at each point. The average registration error is $0.72mm$ on the 100 simulated brain images [70], by calculating the voxelwise difference between the estimated and the ground-truth deformations. We apply our H+DP method to the same simulated dataset and find that the registration error can be reduced to $0.62mm$, which shows the advantage of unsupervised learning of features for guiding registration.

In [3], 14 deformable image registration methods are comprehensively evaluated on various datasets including the LONI data. The best registration method could achieve $71.85 \pm 9.59\%$ as reported in [3]. Our H+DP improves the overlap ratio from $70.47 \pm 10.73\%$ (by HAMMER) to $71.91 \pm 9.57\%$, which is slightly better than the top-ranked registration method in [3].

D. Comparison with Other Feature Learning Methods

There are many unsupervised feature extraction/learning approaches. Here, we further evaluate the power of shallow models (such as PCA and k-means) and our deep learning model. For k-means, we cluster the pre-observed image patches into 128 centers (same as the reduced dimension by PCA) in the feature space. For each new image patch, the membership (i.e., patch-wise distance) to each center forms the feature vector. Table IV shows the registration accuracy on ADNI dataset after integrating the learned feature representations by PCA (H+PCA), k-means (H+KMEAN), and deep learning (H+DP). It can be seen that the registration performance by H+KMEANS is slightly worse than H+PCA. But, as shallow models, both of these PCA and k-means models perform worse than H+DP that use the deep learning model to obtain hierarchical feature.

TABLE IV
THE DICE RATIOS OF WM, GM, AND VN ON ADNI DATASET (UNIT: %)

Method	WM	GM	VN	Overall
H+PCA	86.5	76.9	91.7	85.0
H+KMEAN	86.4	76.9	91.6	85.0
H+DP	88.1	78.6	93.0	86.6

V. Conclusion

A new deformable image registration framework is developed that uses deep learning for feature selection. Specifically, an unsupervised deep learning feature selection framework is proposed that implements a convolutional-stacked auto-encoder network (SAE) to identify the intrinsic features in 3D image patches. Using the LONI and ADNI brain datasets, the image registration performance was compared to two existing state-of-the-art deformable image registration frameworks that use handcrafted features. The results showed the new image

registration framework consistently demonstrated better Dice ratio scores when compared to state-of-the-art. We contribute these increases in performance to our proposed feature selection framework. In short, because the trained deep learning network selected features that more accurately capture the complex morphological patterns in the image patches, this resulted in better anatomical correspondences, which ultimately improved image registration performance.

To demonstrate the scalability of the proposed registration framework, image registration experiments were also conducted on 7.0-tesla brain MR images. Likewise, the results showed the new image registration framework consistently demonstrated better Dice ratio scores when compared to state-of-the-art. Unlike those existing image registration frameworks, the deep learning architecture was quickly developed, trained using no ground-truth data, and still showed superior registration performance. This experiment demonstrates how the proposed feature selection framework can be quickly used to perform image registration on the new imaging modalities.

REFERENCES

[1] W. R. Crum, T. Hartkens, and D. L. G. Hill, "Non-rigid image registration: theory and practice," *British Journal of Radiology*, vol. 77, pp. S140-153, 2004.

[2] K. J. Friston, J. Ashburner, C. D. Frith, J. B. Poline, J. D. Heather, and R. S. J. Frackowiak, "Spatial registration and normalization of images," *Human Brain Mapping*, vol. 3, pp. 165-189, 1995.

[3] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *NeuroImage*, vol. 46, pp. 786-802, 2009.

[4] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *Medical Imaging, IEEE Transactions on*, vol. 21, pp. 1421-1439, 2002.

[5] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977-1000, 2003.

[6] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-euclidean framework for statistics on diffeomorphisms," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, ed, 2006, pp. 924-931.

[7] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, pp. 95-113, 2007.

[8] B. Avants, M. Grossman, and J. Gee, "Symmetric diffeomorphic image registration: evaluating automated labeling of elderly and neurodegenerative cortex and frontal lobe," in *Biomedical Image Registration*, 2006, pp. 50-57.

[9] 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," *Medical Imaging, IEEE Transactions on*, vol. 18, pp. 712-721, 1999.

[10] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: efficient non-parametric image registration," *NeuroImage*, vol. 45, pp. S61-S72, 2009.

[11] Q. Wang, G. Wu, P.-T. Yap, and D. Shen, "Attribute vector guided groupwise registration," *NeuroImage*, vol. 50, pp. 1485-1496, 2010.

[12] T. Rohlfing, "Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable," *IEEE Transactions on Medical Imaging*, vol. 31, pp. 153-160, 2012.

[13] G. Wu, M. Kim, Q. Wang, and D. Shen, "S-HAMMER: Hierarchical Attribute-Guided, Symmetric Diffeomorphic Registration for MR Brain Images," *Human Brain Mapping*, 2013.

[14] G. Wu, P.-T. Yap, M. Kim, and D. Shen, "TPS-HAMMER: Improving HAMMER registration algorithm by soft correspondence matching and thin-plate splines based deformation interpolation," *NeuroImage*, vol. 49, pp. 2225-2233, 2010.

[15] D. Shen and C. Davatzikos, "HAMMER: Hierarchical attribute matching mechanism for elastic registration," *IEEE Transactions on Medical Imaging*, vol. 21, pp. 1421-1439, 2002.

[16] Y. Zhan and D. Shen, "Deformable segmentation of 3-D ultrasound prostate images using statistical texture matching method," *IEEE Transactions on Image Processing*, vol. 25, pp. 256-272, 2006.

[17] G. Wu, Q. Wang, H. Jia, and D. Shen, "Feature-based Groupwise Registration by Hierarchical Anatomical Correspondence Detection," *Human Brain Mapping*, vol. 33, pp. 253-271, 2012.

[18] G. Wu, F. Qi, and D. Shen, "Learning-based deformable registration of MR brain images," *Medical Imaging, IEEE Transactions on*, vol. 25, pp. 1145-1157, 2006.

[19] G. Wu, F. Qi, and D. Shen, "Learning best features and deformation statistics for hierarchical registration of MR brain images," in *Information Processing in Medical Imaging*, ed, 2007, pp. 160-171.

[20] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos, "DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting," *Medical Image Analysis*, vol. 15, pp. 622-639, 2011.

[21] J. Jiang, S. Zheng, A. W. Toga, and Z. Tu, "Learning based coarse-to-fine image registration," presented at the Computer Vision and Pattern Recognition, Anchorage, 2008.

[22] D. Lee, M. Hofmann, F. Steinke, Y. Altun, N. D. Cahill, and B. Scholkopf, "Learning Similarity Measure for Multi-Modal 3D Image Registration," presented at the Computer Vision and Pattern Recognition, Maimi, 2009.

[23] R. W. K. So and A. C. S. Chung, "Learning-based non-rigid image registration using prior joint intensity distributions with graph-cuts," presented at the 18th International Conference on Image Processing, Brussels, 2011.

[24] G. Wu, F. Qi, and D. Shen, "A general learning framework for non-rigid image registration," presented at the Medical Imaging and Augmented Reality, 2006.

[25] M. Kim, G. Wu, W. Li, L. Wang, Y.-D. Son, Z.-H. Cho, *et al.*, "Automatic Hippocampus Segmentation of 7.0 Tesla MR Images by Combining Multiple Atlases and Auto-Context Models," *NeuroImage*, accepted.

[26] P. Comon, " Independent component analysis, a new concept?," *Signal processing*, vol. 36, pp. 287-314, 1994.

[27] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*: John Wiley & Sons, 2001.

[28] J. B. Tenenbaum, V. Silva, and J. C. Langford, " A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 640-646, 2000.

[29] E.-N. I, Y. Y, W. MN, G. NP, and N. RM, "A support vector machine approach for detection of microcalcifications," *IEEE Transaction on Medical Imaging*, vol. 21, pp. 1552-1563, 2002.

[30] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," presented at the Proceedings of the 24th international conference on Machine learning, Corvalis, Oregon, 2007.

[31] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, p. 2323, 2000.

[32] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373-1396, 2003.

[33] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607-609, 1996.

[34] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *Arxiv*, 2012.

[35] L. Arnold, S. Rebecchi, S. Chevallier, and H. Paugam-moisy, "An introduction to deep-learning," presented at the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2011.

[36] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," presented at the Advances in Neural Information Processing Systems (NIPS), 2006.

[37] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527-1554, 2006.

[38] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313,

pp. 504-507, 2006.

[39] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis " presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011.

[40] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," presented at the Advances in Neural Information Processing Systems (NIPS), 2008.

[41] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," presented at the Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, 2009.

[42] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," *Commun. ACM*, vol. 54, pp. 95-103, 2011.

[43] Y. LeCun and Y. Bengio, "Convolutional network for images, speach, and time series," in *The handbook of brain theory and neural networks*, ed, 1995.

[44] H.-C. Shin, M. Orton, D. Collins, S. Doran, and M. Leach, "Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1930-1943, 2013.

[45] G. Wu, M. Kim, Q. Wang, S. Liao, Y. Gao, and D. Shen, "Unsupervised Deep Feature Learning for Deformable Image Registration of MR Brains," presented at the MICCAI 2013, Nagoya, Japan, 2013.

[46] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, *et al.*, "Construction of a 3D probabilistic atlas of human cortical structures," *NeuroImage*, vol. 39, pp. 1064-1080, 2008.

[47] J. MacQueen, "Some methods for classification and analysis of mulitvariate observations," *In Proceeding of the fifth Berkeley symposium on mathematical statistics and probability* vol. 1, pp. 281-297, 1967.

[48] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*: John Wiley & Sons, 1986.

[49] A. Coates and A. Y. Ng, "Learning Feature Representations with K-means," presented at the In Neural Networks: Tricks of the Trade, Reloaded, 2012.

[50] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "kernel PCA and Denoising in Feature Space," *Advances in Neural Information Processing System*, vol. 11, pp. 536-542, 1999.

[51] T. F. Cootes, D. Cooper, C. J. Taylor, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 16, pp. 38-59, 1995.

[52] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 681-685, 2001.

[53] M. A. Ranzato, C. Poultney, S. Chopra, and Y. Lecun, "Efficient learning of sparse representations with an energy-based model," presented at the Advances in Neural Information Processing Systems (NIPS), 2006.

[54] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On Optimization Methods for Deep Learning," presented at the ICML, 2011.

[55] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, pp. 83-105, 2001.

[56] X. Pennec, P. Cachier, and N. Ayache, "Understanding the "Demon's Algorithm": 3D Non-rigid Registration by Gradient Descent," presented at the Medical Image Computing and Computer-Assisted Intervention, Cambridge, UK, 1999.

[57] D. Shen, "Image registration by local histogram matching," *Pattern Recognition*, vol. 40, pp. 1161-1172, 2007.

[58] J. M. Peyrat, H. Delingette, M. Sermesant, X. Pennec, C. Xu, and N. Ayache, "Registration of 4D time-series of cardiac images with multichannel Diffeomorphic Demons," presented at the Med Image Comput Comput Assist Interv, 2008.

[59] J. M. Peyrat, H. Delingette, M. Sermesant, C. Xu, and N. Ayache, "Registration of 4D cardiac CT sequences under trajectory constraints with multichannel diffeomorphic demons," *IEEE Trans Med Imaging*, vol. 29, pp. 1351-68, Jul 2010.

[60] D. Forsberg, Y. Rathi, S. Bouix, D. Wassermann, H. Knutsson, and C.-F. Westin, "Improving Registration Using Multi-channel Diffeomorphic Demons Combined with Certainty Maps," presented at the Multimodal Brain Image Registration, Lecture Notes in Computer Science Volume 7012, 2011.

[61] D. G. Shen, "Image registration by local histogram matching," *Pattern Recognition*, vol. 40, pp. 1161-1172, Apr 2007.

[62] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, pp. 567-585, 1989.

[63] F. Shi, L. Wang, Y. Dai, J. Gilmore, W. Lin, and D. Shen, "LABEL: Pediatric Brain Extraction using Learning-based Meta-algorithm," *NeuroImage*, vol. 62, pp. 1975-1986, 2012.

[64] N. Tustison, B. Avants, P. Cook, Y. Zheng, A. Egan, P. Yushkevich, *et al.*, "N4ITK: Improved N3 Bias Correction," *IEEE Trans Medical Imaging*, vol. 29, pp. 1310-1320, 2010.

[65] A. Madabhushi and J. Udupa, "New methods of MR image intensity standardization via generalized scale," *Medical Physics*, vol. 33, pp. 3426-3434, 2006.

[66] G. E. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," in *Neural Networks: Tricks and Trade*. vol. 7700, ed: LNCS, 2012, pp. 599-619.

[67] M. M. Shattuck DW, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, Poldrack RA, Bilder RM, Toga AW, "Construction of a 3D Probabilistic Atlas of Human Cortical Structures," *NeuroImage*, vol. 39, pp. 1064-1080, 2008.

[68] Z.-H. Cho, J.-Y. Han, S.-I. Hwang, D.-s. Kim, K.-N. Kim, N.-B. Kim, *et al.*, "Quantitative analysis of the hippocampus using images obtained from 7.0 T MRI," *NeuroImage*, vol. 49, pp. 2134-2140, 2010.

[69] Z.-H. Cho, Y.-B. Kim, J.-Y. Han, H.-K. Min, K.-N. Kim, S.-H. Choi, *et al.*, "New brain atlas—Mapping the human brain in vivo with 7.0 T MRI and comparison with postmortem histology: Will these images change modern medicine?," *International Journal of Imaging Systems and Technology*, vol. 18, pp. 2-8, 2008.

[70] Z. Xue, D. Shen, B. Karacali, J. Stern, D. Rottenberg, and C. Davatzikos, "Simulating deformations of MR brain images for validation of atlas-based segmentation and registration algorithms," *NeuroImage*, vol. 33, pp. 855-866, 2006.