Neuroinformatics (2020) 18:319–331
https://doi.org/10.1007/s12021-019-09448-5

ORIGINAL ARTICLE

FCN Based Label Correction for Multi-Atlas Guided Organ Segmentation

Hancan Zhu[1] · Ehsan Adeli[2] · Feng Shi[3] · Dinggang Shen[4,5] · for the Alzheimer's Disease Neuroimaging Initiative

Published online: 2 January 2020

Abstract

Segmentation of medical images using multiple atlases has recently gained immense attention due to their augmented robustness against variabilities across different subjects. These atlas-based methods typically comprise of three steps: atlas selection, image registration, and finally label fusion. Image registration is one of the core steps in this process, accuracy of which directly affects the final labeling performance. However, due to inter-subject anatomical variations, registration errors are inevitable. The aim of this paper is to develop a deep learning-based confidence estimation method to alleviate the potential effects of registration errors. We first propose a fully convolutional network (FCN) with residual connections to learn the relationship between the image patch pair (i.e., patches from the target subject and the atlas) and the related label confidence patch. With the obtained label confidence patch, we can identify the potential errors in the warped atlas labels and correct them. Then, we use two label fusion methods to fuse the corrected atlas labels. The proposed methods are validated on a publicly available dataset for hippocampus segmentation. Experimental results demonstrate that our proposed methods outperform the state-of-the-art segmentation methods.

Keywords Multi-atlas image segmentation · Label fusion · Fully convolutional network · Deep learning

Introduction

Recently, multi-atlas image segmentation (MAIS) methods have become increasingly the most reliable methods for medical image segmentation (Iglesias and Sabuncu 2015). MAIS exploits several labeled atlases to segment a target image; each atlas consists of an image and its corresponding label map, which is usually obtained by manual segmentation. These algorithms often comprise three steps: atlas selection, image registration, and label fusion. Specifically, the first step is to select most relevant atlases (based on similarity indices). Then, the selected atlases are registered to the target image (i.e., the image to be segmented), and the corresponding atlas labels are warped to the target image space with the obtained registration deformation fields. Finally, the warped atlas labels are fused to obtain the final segmentation (referred to as label fusion). In this process, accurate registrations between atlas images and the target image is crucial. However, image registration is an ill-posed problem given the large inter-subject anatomical variances (Haber and Modersitzki 2004). Atlas images cannot be perfectly matched to the target image, and thus registration errors are inevitable (Doshi et al. 2016). The majority of research works on MAIS utilize existing registration tools but focus more on the atlas selection and label fusion steps to counter the registration errors (Doshi et al. 2016; Aljabar et al. 2009; Artaechevarria et al. 2009; Asman and Landman 2013; Benkarim et al. 2017; Zhu et al. 2017; Hao et al. 2014).

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

✉ Dinggang Shen
dgshen@med.unc.edu

[1] School of Mathematics Physics and Information, Shaoxing University, Shaoxing 312000, China

[2] Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford 94305, CA, USA

[3] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China

[4] Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill 27599, North Carolina, USA

[5] Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea

The existing atlas selection (Aljabar et al. 2009; Hao et al. 2014; Rohlfing et al. 2004; Cao et al. 2011; Duc et al. 2012; Langerak et al. 2013; Lötjönen et al. 2010; Sanroma et al. 2014; Zaffino et al. 2018) and label fusion (Artaechevarria et al. 2009; Asman and Landman 2013; Hao et al. 2014; Rohlfing et al. 2004; Heckemann et al. 2006; Coupé et al. 2011; Rousseau et al. 2011; Liao et al. 2012; Wang et al. 2013; Wang et al. 2011; Bai et al. 2015; Zhu et al. 2015; Warfield et al. 2004; Bai et al. 2013; Jorge Cardoso et al. 2013; Sabuncu et al. 2010; Liao et al. 2013; Haom et al. 2012; Asman and Landman 2012; Commowick et al. 2012; Asman and Landman 2014) methods have shown their effectiveness in alleviating registration errors in the MAIS methods; however, the most natural way to address the problem is to find the potential errors in the warped atlas labels and then correct them before performing label fusion. In this paper, we propose a deep learning-based confidence estimation method for detecting the potential errors in the warped atlas labels. Those detected labels in the warped atlas are then corrected, and two label fusion schemes are used to fuse the corrected labels to obtain the final segmentation. Figure 1 shows the general framework of the proposed method. We validate the proposed methods for hippocampus segmentation using a publicly available dataset (Boccardi et al. 2015). The proposed methods are compared with several state-of-the-art segmentation methods, including majority voting (MV) (Rohlfing et al. 2004; Heckemann et al. 2006), joint label fusion (JLF) (Wang et al. 2013), and a deep learning segmentation method with 3D deeply supervised network (DSN) (Dou et al. 2017). The obtained results show that the proposed methods outperform the state-of-the-art methods in terms of several segmentation evaluation metrics.

The main contributions of this work can be summarized as follows: 1) We propose a novel multi-atlas image segmentation framework by estimating the confidence of each warped atlas label, used to identify and correct all warped atlas labels; 2) Instead of using local supervised learning models, we propose a deep learning based global model to learn the relationship between the image patch pairs (the target image patch and the atlas image patch) and the confidence of each warped label in the atlas patch; 3) The proposed method can combine the advantages of multi-atlas segmentation method and deep learning based segmentation methods to improve the segmentation accuracy and robustness.

Background and Related Work

The atlas selection step selects a subset of atlases that are most similar to the target image based on certain image similarity criteria, such as normalized mutual information (Aljabar et al. 2009; Hao et al. 2014; Rohlfing et al. 2004), distance in lower dimensional manifold space (Cao et al. 2011; Duc et al. 2012), and registration performance (Langerak et al. 2013). Since dissimilar atlases may produce more severe registration errors when registered to the target image, removing them can intuitively improve MAIS performance. However, a main issue is that image similarity metrics cannot always guarantee the optimal selection of atlases (Lötjönen et al. 2010; Sanroma et al. 2014) due to various inter-subject variability. To make the methods agnostic to similarity measures and their induced bias, learning based methods (such as support vector machine-based atlas ranking (Sanroma et al. 2014)) has been used for atlas selection. In a recent paper (Zaffino et al. 2018), Zaffino et al. argued that one should prefer the best atlas group over the group of best atlases, and subsequently proposed an atlas group selection algorithm based on convolutional neural networks.

In the label fusion step, the warped atlases are combined to obtain the final segmentation (Iglesias and Sabuncu 2015). The existing label fusion methods can be mainly categorized into three different categories: voting methods (Artaechevarria et al. 2009; Rohlfing et al. 2004; Heckemann et al. 2006; Coupé et al. 2011; Rousseau et al. 2011; Liao et al. 2012; Wang et al. 2013), learning-based methods (Hao et al. 2014; Wang et al. 2011; Bai et al. 2015; Zhu et al. 2015), and

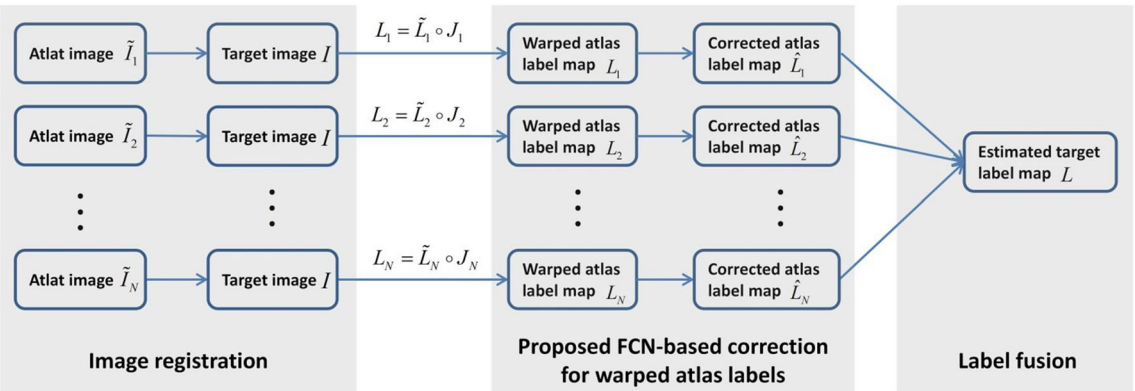

Fig. 1 The general framework of the proposed method. In the figure, L̃_i is the i-th atlas label map, J_i is the deformation field obtained by registering the i-th atlas image to the target image

probabilistic methods (Asman and Landman 2013; Warfield et al. 2004; Bai et al. 2013; Jorge Cardoso et al. 2013). Majority voting is the simplest voting method, which assigns the same weighting coefficient to each atlas (Rohlfing et al. 2004; Heckemann et al. 2006). The voting label fusion methods use a weighted combination of atlas labels to obtain the target image segmentation. Different combination strategies have been investigated, including global weighted voting (Sabuncu et al. 2010) and local weighted voting (Artaechevarria et al. 2009). It is shown that local weighed voting scheme outperforms global weighted voting when segmenting high-contrast structures, but global methods are less sensitive to noise when segmenting low-contrast structures (Artaechevarria et al. 2009). To further relieve the registration errors, non-local weighted voting methods have been proposed in the literature, which select training samples in a searching neighborhood from each atlas for voting (Coupé et al. 2011; Rousseau et al. 2011). Sparse representation and joint label fusion methods have been also widely investigated for improving the weighted voting label fusion results (Wang et al. 2013; Liao et al. 2013). In the learning-based methods, a local regression or classification model is built to model the relationship between the image appearances and the corresponding labels using the samples obtained from the neighboring region of each atlas (Wang et al. 2011; Haom et al. 2012). To obtain better segmentation results, augmented features are usually used in these methods. For example, the first- and second-order gradient filters, Sobel and Laplacian operators are used in (Hao et al. 2014), while image gradients, context features and image intensities are used in (Bai et al. 2015). As an example for probabilistic methods, Bayesian approaches were used for label fusion (Warfield et al. 2004). In (Warfield et al. 2004), the STAPLE algorithm was introduced to iteratively estimate the target segmentation and the performance of each atlas. Several methods have been proposed to improve the STAPLE method, including the local STAPLE by estimating the reference segmentation with spatially varying performance parameters (Asman and Landman 2012; Commowick et al. 2012), non-local STAPLE by reformulating the STAPLE framework from a non-local mean perspective (Asman and Landman 2013), and hierarchical STAPLE using hierarchical models of rater performance (Asman and Landman 2014).

The multi-atlas image segmentation method (MAIS) (Iglesias and Sabuncu 2015) utilizes N selected atlases $\widetilde{A}_i = \left(\widetilde{I}_i, \widetilde{L}_i\right), i = 1, 2, \ldots, N$ to segment a target image I. For the i-th atlas $\widetilde{A}_i$, let $\widetilde{I}_i$ be its atlas image and $\widetilde{L}_i$ the corresponding label map. In MAIS, each atlas image is first registered to the target image and its corresponding label map is then propagated to the target image space, resulting N warped atlases, i.e., $A_i = (I_i, L_i), \forall\, i = 1, 2, \ldots, N$. Then, the label of each voxel in the target image is inferred from the warped atlases. This procedure is referred to as label fusion.

One of the most widely used label fusion methods is the weighted voting scheme, in which the label of the target voxel x is computed by a weighted combination of the corresponding warped atlas labels (Artaechevarria et al. 2009),

$$L(x) = \underset{l}{\operatorname{argmax}} \sum_{i=1}^{N} w_i(x)(L_i(x) == l), \tag{1}$$

where $w_i(x)$ is the weight of the i-th atlas voxel x reflecting the confidence for the i-th atlas image. The simplest way to set these weights in (1) is to use constant weights for all atlases, leading to the majority voting label fusion method (Rohlfing et al. 2004; Heckemann et al. 2006). However, the proper estimation of weights $w_i(x)\ \forall\, i = 1, 2, \ldots, N$ improves the overall segmentation performance. Global weighted voting methods (e.g., (Sabuncu et al. 2010; Artaechevarria et al. 2008)) estimate a global weight w_i for each atlas, irrespective of the voxel location. But given the fact that the registration error may be distributed differently at different locations in each atlas, estimating local weight $w_i(x)$ for each atlas at each voxel location may be more feasible. The weight is usually estimated according to the local appearance and similarity between the target image and each atlas, measured by a similarity function such as Gaussian function (Sabuncu et al. 2010),

$$w_i(x) = e^{\frac{-\|p_t(x) - p_i(x)\|_2^2}{\sigma_x}},$$

where $p_t(x)$ and $p_i(x)$ are the same size vectorized patches centered at x on the target and the i-th atlas image, respectively; and σ_x is a tuning hyperparameter.

To further alleviate possible registration errors, non-local weighted voting (NLW) methods were proposed (Coupé et al. 2011; Rousseau et al. 2011), in which the label of the target voxel can be computed by

$$L(x) = \underset{l}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{y \in \Omega_x} w_i(x, y)(L_i(y) == l), \tag{2}$$

where Ω_x is a search region centered at voxel x, and $w_i(x, y)$ is the weight reflecting the confidence that the voxel y of the i-th atlas has the same segmentation label as the target voxel x. The weight $w_i(x, y)$ in (2) is estimated by the local appearance similarity between patches of the target voxel x and the voxel y in the i-th atlas image, according to the Gaussian similarity function (Coupé et al. 2011; Rousseau et al. 2011),

$$w_i(x, y) = e^{\frac{-\|p_t(x) - p_i(y)\|_2^2}{\sigma_x}}.$$

In other works, based on these NLW label fusion methods, Wang et al. (Wang et al. 2013) proposed a joint probability model for estimating the confidence weights, Liao et al. (Liao et al. 2013) proposed a sparse representation method for

estimating the confidence weights and Zhu et al. (Zhu et al. 2017) proposed a local supervised metric learning method for estimating the confidence weights, which were then used to compute the target label through weighted voting.

Recently, with the surge of deep learning methods, such technologies have also been applied to the tasks of multi-atlas label fusion (Yang et al. 2017; Fang et al. 2017), with the advantage of being independent from manual feature extraction schemes. Specifically, in (Yang et al. 2017), the authors formulated multi-atlas segmentation in a deep learning framework, by integrating the feature extraction and the non-local patch-based label fusion in a single deep network architecture. Fang et al. (Fang et al. 2017) introduced a multi-atlas guided FCN by incorporating atlas information within the network learning. Different from these methods, we utilize a deep learning approach to estimate the confidence, which plays a key role in the weighted label fusion methods. With the estimated confidence, the warped atlas labels can be corrected, and then the target label is computed by two label fusion schemes.

Methods

Our proposed deep learning method for multi-atlas label fusion is comprised of a novel fully convolutional network (FCN) to model the relationship between the image patch pair and the label confidence. After learning the confidence, similar to the previous works, two label fusion schemes including majority voting (MV) and joint label fusion (JLF) are used to fuse the corrected warped atlas labels. Therefore, we refer to our methods as FCN-MV and FCN-JLF throughout the paper.

FCN Based Confidence Estimation

Given an image patch p_t in the target image and the corresponding image patch p_a in a warped atlas image, we learn a function $f(p_t, p_a)$ to model the relationship between the image patch pair (p_t, p_a) and the label confidence C (Fig. 2 illustrates this process),

$$C = f(p_t, p_a),$$

where C is a patch with the same size as p_t and p_a, indicating whether p_t and p_a have the same segmentation label,

$$C(x) = \delta(l_t(x), l_a(x)) = \begin{cases} 1, & if\ l_t(x) = l_a(x), \\ 0, & if\ l_t(x) \neq l_a(x), \end{cases}$$

where $l_t(x)$ is the label of the target image voxel $p_t(x)$, and $l_a(x)$ is the label of the warped atlas image voxel $p_a(x)$; hence C can serve as the confidence map. Here, we propose a fully convolutional network (FCN) (Long et al. 2015; Ronneberger et al. 2015) to predict the confidence estimation function $f(p_t, p_a)$ (Milletari et al. 2016; Yu et al. 2017), but unlike regular FCNs, we propose one with residual connections, which has been demonstrated to be effective for promoting information propagation and accelerating the convergence (He et al. 2016a). We incorporate an architecture similar to U-Net (Ronneberger et al. 2015) and hence we refer to our proposed FCN architecture as ResUNet.

Figure 3 shows the structure of ResUNet, which consists of a down-sampling path and an up-sampling path. The down-sampling path contains one $3 \times 3 \times 3$ convolution layer, two $2 \times 2 \times 2$ max-pooling operations with stride 2, and three residual blocks. Correspondingly, the up-sampling path contains three residual blocks, two $4 \times 4 \times 4$ deconvolution layers with stride 2, and one $1 \times 1 \times 1$ convolution layer. Each $3 \times 3 \times 3$ convolution is followed by a batch normalization and a rectified linear unit (ReLU). To retain the spatial and localization details in the up-sampling pathway, padded convolution layers are used in our network, in which feature maps in the down-sampling path are connected to the corresponding features in the up-sampling path through element-wise summation. These long-skip connections can provide detailed image

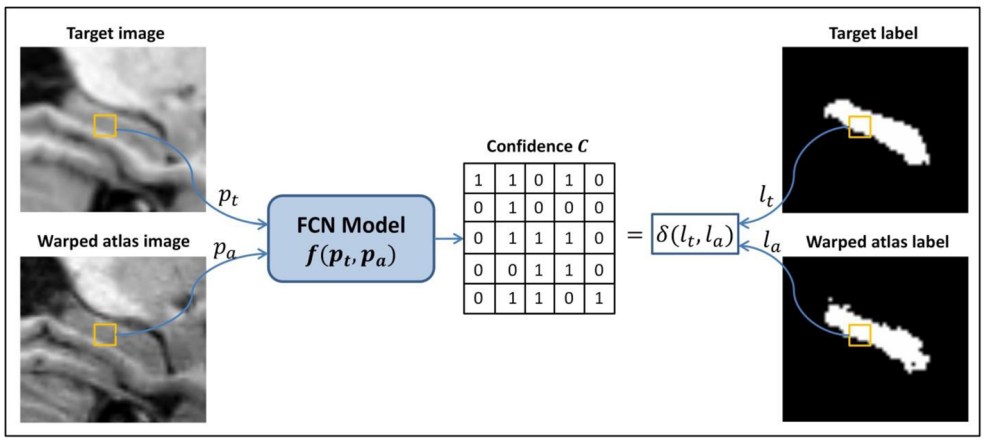

Fig. 2 The illustration of FCN based confidence learning

information to the up-sampling path that is otherwise lost during the successive down-sampling process. The number of possible outputs k, in the last $1 \times 1 \times 1$ convolution layer define the number of classes, which is set to 2 in our application (i.e., we have two classes, with '0' representing different labels and '1' representing the same labels).

The residual block consists of two $3 \times 3 \times 3$ convolutions, each followed by a batch normalization layer and a ReLU. In the residual block, residual connections are used to connect the input features to the output feature maps of last convolution with an element-wise summation operation. Formally, the residual block can be expressed as (He et al. 2016b),

$$\eta = \varphi(\xi) + \xi,$$

where ξ denotes the input feature maps, η denotes the output feature maps, and $\varphi(\cdot)$ is the residual function which consists of two $3 \times 3 \times 3$ convolutions, each followed by a batch normalization layer and a ReLU in our network. As studied previously, the residual connections alleviate the problem of gradient vanishing, promote information propagation, and accelerate the convergence (He et al. 2016a).

To train the model, a softmax loss is used (Gu et al. 2017):

$$L_{Softmax} = -\sum_{i=1}^{m}\sum_{j=0}^{1} 1\{C(x_i) = j\}\log\frac{e^{z_{j,i}}}{\sum_{h=0}^{1} e^{z_{h,i}}},$$

where $z_{j,i}$ represents the j-th output of the last network layer for the i-th voxel, $C(x_i) \in \{0, 1\}$ represents the ground-truth confidence at the location of voxel x_i, and m is the number of voxels in the input patch.

Label Fusion with FCN-Based Confidence Estimation

For labeling the target patch p_t, the corresponding atlas image patch p_i and atlas label patch l_i are extracted from the i-th warped atlas, $i = 1, 2, \ldots, N$. With the trained confidence estimation model, we compute the confidence $C_i = f(p_t, p_i)$ for each patch pair (p_t, p_i), $i = 1, 2, \ldots, N$. Then, we correct label values in each label patch l_i according to the obtained confidence C_i. For the case of binary segmentation (as in our application), we have only two segmentation labels denoted by {0,1}. The corrected label patch $\hat{l}_i$ is, therefore, computed by

$$\hat{l}_i(x) = \begin{cases} l_i(x), & if\ C_i(x) = 1; \\ 1 - l_i(x), & if\ C_i(x) = 0. \end{cases}$$

After label correction, we use two label fusion methods to compute the label values of the target patch, including majority voting (Rohlfing et al. 2004; Heckemann et al. 2006) and joint label fusion (Wang et al. 2013).

With the majority voting label fusion, the target label patch l_t is determined by

$$l_t(x) = \underset{l}{\operatorname{argmax}} \sum_{i=1}^{N} \left(\hat{l}_i(x) == l\right), \quad l \in \{0, 1\}.$$

With the joint label fusion, the target label patch l_t can be computed by

$$l_t(x) = \underset{l}{\operatorname{argmax}} \sum_{i=1}^{N} w_i(\xi_i(x))\left(\hat{l}_i(\xi_i(x)) == l\right), \quad l \in \{0, 1\},$$

where $\xi_i(x)$ is the local search correspondence map between the ith atlas and the target image, and $w_i(\xi_i(x))$ is the weight for the ith atlas. We denote $\vec{w}_x = [w_1(\xi_1(x)); w_2(\xi_2(x)); \ldots; w_N(\xi_N(x))]$. Then, $\vec{w}_x$ is determined by

$$\underset{\vec{w}_x}{\operatorname{argmin}}\ \vec{w}_x^{t}(M_x + \alpha I)\vec{w}_x,$$

$$\text{s.t.}\ \sum_{i=1}^{N} w_i(\xi_i(x)) = 1,$$

where t stands for transpose, I is an identity matrix, α is a parameter ($\alpha = 0.1$), and M_x is a pairwise dependency matrix (Wang et al. 2013).

As we use a patch-wise label fusion, for each target voxel patch, a different label is computed, instead of only taking the center voxel as the representative. The majority voting strategy can hence be used to determine the labels of the overlapping voxels of neighboring patches.

Evaluation Metrics

The image segmentation results are comprehensively evaluated based on nine different segmentation evaluation measures, including Dice coefficient, Jaccard index, Precision, Recall, Mean distance (MD), Hausdorff distance (HD), Hausdorff 95 distance (HD95), Average Symmetric Surface Distance (ASSD), and Root Mean Square Distance (RMSD) (Jafari-Khouzani et al. 2011). The first four metrics are used to measure the relative volumetric overlap between the automated segmentation and the ground-truth segmentation, and the last five metrics were used to measure the agreement between segmentation boundaries. By denoting A as the manual segmentation, B as the automated segmentation, and $V(X)$ as the volume of segmentation X, these nine evaluation measures can be defined as:

$$\text{Dice} = 2\frac{\text{V(A}\cap\text{B)}}{\text{V(A)} + \text{V(B)}}, \text{Jaccard} = \frac{\text{V(A}\cap\text{B)}}{\text{V(A}\cup\text{B)}}, \text{Precision} = \frac{\text{V(A}\cap\text{B)}}{\text{V(B)}},$$

$$\text{Recall} = \frac{\text{V(A}\cap\text{B)}}{\text{V(A)}}, \text{MD} = \text{mean}_{e\in\partial A}\left(min_{f\in\partial B} d(e, f)\right),$$

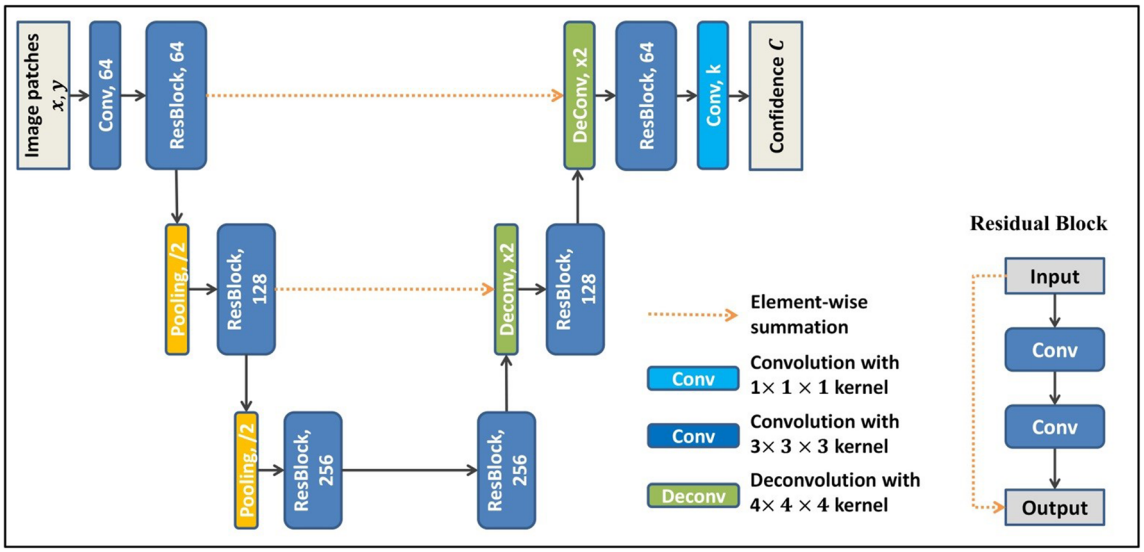

Fig. 3 The illustration of the proposed ResUNet structure. The number of kernels is denoted in each convolution operation rectangle

$$\mathrm{HD} = \max(\mathrm{H}(\mathrm{A},\mathrm{B}),\mathrm{H}(\mathrm{B},\mathrm{A})), \quad \text{where } \mathrm{H}(\mathrm{A},\mathrm{B}) = max_{e\in\partial A}\left(min_{f\in\partial B}d(e,f)\right),$$

HD95:similar to HD,except that 5% data points with the largest distance are removed before calculation,

$$\mathrm{ASSD} = \left(\mathrm{mean}_{e\in\partial A}\left(min_{f\in\partial B}d(e,f)\right) + \mathrm{mean}_{e\in\partial B}\left(min_{f\in\partial A}d(e,f)\right)\right)/2,$$

$$\mathrm{RMSD} = \frac{\sqrt{D_A^2 + D_B^2}}{card\{\partial A\} + card\{\partial B\}}, \quad \text{where } D_A^2 = \sum_{e\in\partial A}\left(min_{f\in\partial B}d(e,f)\right),$$

where ∂A is the boundary voxels of A, $d(\cdot,\cdot)$ is the Euclidian distance of two points, and $card\{\cdot\}$ is the cardinality of a set.

Experiments and Results

Data and Preprocessing

The proposed method is validated for hippocampus segmentation using a subset of the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/), containing 100 T1 MR images (29 normal controls, 34 subjects with mild cognitive impairment, and 37 subjects with Alzheimer's disease). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For detailed information, see www.adni-info.org. The ADNI MRI scans were acquired using a sagittal 3D MP-RAGE T1-w sequence (TR = 2400 ms, minimum full TE, TI = 1000 ms, FOV = 240 mm, voxel size of 1.25 × 1.25 × 1.2mm^3) (Jack et al. 2008). Ground-truth hippocampus labels of the image data were provided in a preliminary release part by the EADC–ADNI (European Alzheimer's Disease Consortium and Alzheimer's Disease Neuroimaging Initiative) harmonized segmentation protocol (www.hippocampal-protocol.net) (Boccardi et al. 2015). All MR images were aligned along the line passing through the anterior and posterior commissures of the brain (AC-PC line), and their bias fields were corrected. Then, all images were spatially normalized to the MNI152 template with the voxel size of 1 × 1 × 1 mm^3, using affine transformation (Boccardi et al. 2015).

In our experiments, we randomly select 40 subjects as atlases. For the remaining 60 subjects, a two-fold cross-validation strategy is used to evaluate the segmentation performance. Specifically, we randomly divide the 60 subjects into two partitions (i.e., folds), each containing 30 subjects. For each fold, one partition is used for training the confidence estimation model using our proposed ResUNet, and the other for testing the model performance. During training, 3 of 30 subjects were randomly selected for validation.

To reduce the computational cost, we run the algorithm on the cropped hippocampus box, identified by a simple preprocessing step. Since all the images were linearly aligned to the MNI152 template, we scan all training atlases to find the minimum and maximum positions of the left and right hippocampi along the axial, coronal, and sagittal directions, and then enlarge the obtained box by 7 voxels in each direction to form the cropping boxes for the left and right hippocampi, respectively, thus they are big enough to cover the hippocampi of unseen testing subjects. All images are then cropped using

Table 1 Dice values (mean±std) of hippocampus segmentation results using FCN-MV with different patch sizes ($r_p \times r_p \times r_p$)

	$r_p = 4$	$r_p = 8$	$r_p = 12$	$r_p = 16$
Left	0.880 ± 0.023	0.883 ± 0.022	0.883 ± 0.021	0.880 ± 0.023
Right	0.886 ± 0.023	0.888 ± 0.021	0.886 ± 0.022	0.881 ± 0.025

these identified boxes, and the cropped images are normalized to have similar intensity levels by using a histogram matching method. A nonlinear, cross-correlation-driven image registration algorithm (Avants et al. 2008) is used to register the cropped atlas images to each cropped target image.

Experiment and Parameter Setting

To further reduce the computational cost, we use the majority voting label fusion method in (Heckemann et al. 2006) to obtain an initial segmentation of the target image. We, then, apply the proposed method only to the voxels without 100% votes for either the hippocampus or the background in the majority voting method. We randomly extract image patch pairs from each training image and its warped atlas images centered at the locations where 100% votes were not achieved in the majority-voting based initial segmentation. The patch size is empirically selected from $4 \times 4 \times 4$, $8 \times 8 \times 8$, $12 \times 12 \times 12$ and $16 \times 16 \times 16$. Table 1 shows the segmentation results obtained by FCN-MV with different patch sizes. It can be observed that the results obtained by FCN-MV with the patch size $8 \times 8 \times 8$ are slightly better than those with other 3 patch sizes. Thus, we set the patch size to $8 \times 8 \times 8$ in the proposed methods, FCN-MV and FCN-JLF.

Our network is trained on a NVIDIA Titan Xp with 12 GB memory, and implemented using Caffe (Jia et al. 2014). The Adam optimizer is used for training with a batch size of 20. The learning rate is initially set to 0.0001 and then decreased by a factor of $\gamma = 0.1$ every 10,000 iterations. The weight decay and momentum are set to 0.0005 and 0.9, respectively. The network is trained for maximum 60,000 iterations. Note that we separately construct the training set and then train separate ResUNet models for the left and right hippocampi, respectively.

Comparison with Existing Methods

We compare our proposed methods, FCN-MV and FCN-JLF, with two widely-used label fusion methods, MV (Rohlfing et al. 2004; Heckemann et al. 2006) and JLF (Wang et al. 2013), and also with a deep learning segmentation method with 3D deeply supervised network (DSN) (Dou et al. 2017). The two label fusion methods are running the same settings as our proposed methods (i.e., the same set of 40 atlases, same non-linear registration, and same patch-wise label fusion fashion). Atlas selection is conducted based on normalized mutual information (NMI) for selecting the top 20 most similar atlases from the atlas set (Zhu et al. 2017; Hao et al. 2014). The optimal hyperparameters of JLF were $r_p = 1$ and $\beta = 1$, which are selected from {1, 2, 3} and {0.5, 1, 1.5, 2}, using a grid-search strategy based on the atlas dataset with 40 leave-one-out cross-validation experiments.

Table 2 Nine index values (mean±std) of hippocampus segmentation results using different methods

	MV	JLF	DSN	FCN-MV	FCN-JLF
Dice	0.856 ± 0.031#*	0.880 ± 0.024#*	0.869 ± 0.022#*	0.883 ± 0.022	**0.884 ± 0.020**
(L/R)	0.860 ± 0.033#*	0.884 ± 0.023#*	0.871 ± 0.024#*	0.888 ± 0.021*	**0.891 ± 0.019**
Jaccard	0.750 ± 0.047#*	0.786 ± 0.037#*	0.769 ± 0.035#*	0.792 ± 0.034	**0.793 ± 0.032**
(L/R)	0.755 ± 0.048#*	0.794 ± 0.036#*	0.772 ± 0.036#*	0.800 ± 0.033*	**0.803 ± 0.030**
Precision	0.861 ± 0.048#*	0.879 ± 0.032#	0.866 ± 0.033#*	**0.896 ± 0.029**	0.879 ± 0.029#
(L/R)	0.864 ± 0.052#*	0.882 ± 0.036#*	0.870 ± 0.034#*	**0.902 ± 0.033**	0.889 ± 0.0.031#
Recall	0.854 ± 0.049#*	0.882 ± 0.036#	0.874 ± 0.032*	0.872 ± 0.033*	**0.890 ± 0.030**
(L/R)	0.859 ± 0.044#*	0.889 ± 0.029#	0.873 ± 0.034*	0.876 ± 0.028*	**0.894 ± 0.027**
HD	3.157 ± 0.853#*	3.076 ± 0.784#	7.324 ± 10.672#*	**2.843 ± 0.770**	2.951 ± 0.827
(L/R)	3.255 ± 0.894#*	3.227 ± 1.100	5.250 ± 8.067	**3.013 ± 0.878**	3.057 ± 0.993
HD95	1.345 ± 0.478#*	1.093 ± 0.352	1.934 ± 4.532#*	1.054 ± 0.271	**1.028 ± 0.167**
(L/R)	1.332 ± 0.441#*	1.101 ± 0.237	1.322 ± 0.294#*	1.099 ± 0.244	**1.070 ± 0.196**
MD	0.284 ± 0.054#*	0.252 ± 0.048#*	0.317 ± 0.296#*	**0.218 ± 0.032**	0.238 ± 0.033#
(L/R)	0.278 ± 0.063#*	0.237 ± 0.051#*	0.258 ± 0.058#*	**0.205 ± 0.044**	0.221 ± 0.043#
ASSD	0.334 ± 0.077#*	0.265 ± 0.052*	0.353 ± 0.145#*	0.255 ± 0.042	**0.252 ± 0.035**
(L/R)	0.328 ± 0.071#*	0.260 ± 0.043#*	0.322 ± 0.049#*	0.249 ± 0.036*	**0.242 ± 0.034**
RMSD	0.632 ± 0.123#*	0.551 ± 0.090#*	0.874 ± 0.910#*	0.533 ± 0.073	**0.527 ± 0.058**
(L/R)	0.628 ± 0.110#*	0.550 ± 0.074#*	0.679 ± 0.293#*	0.534 ± 0.067*	**0.523 ± 0.063**

'#' indicates FCN-MV achieves significant improvement over the corresponding method and '*' indicates FCN-JLF achieves significant improvement over the corresponding method in the Wilcoxon signed rank tests with p value <0.05. Best results in each row are typeset in bold typeface

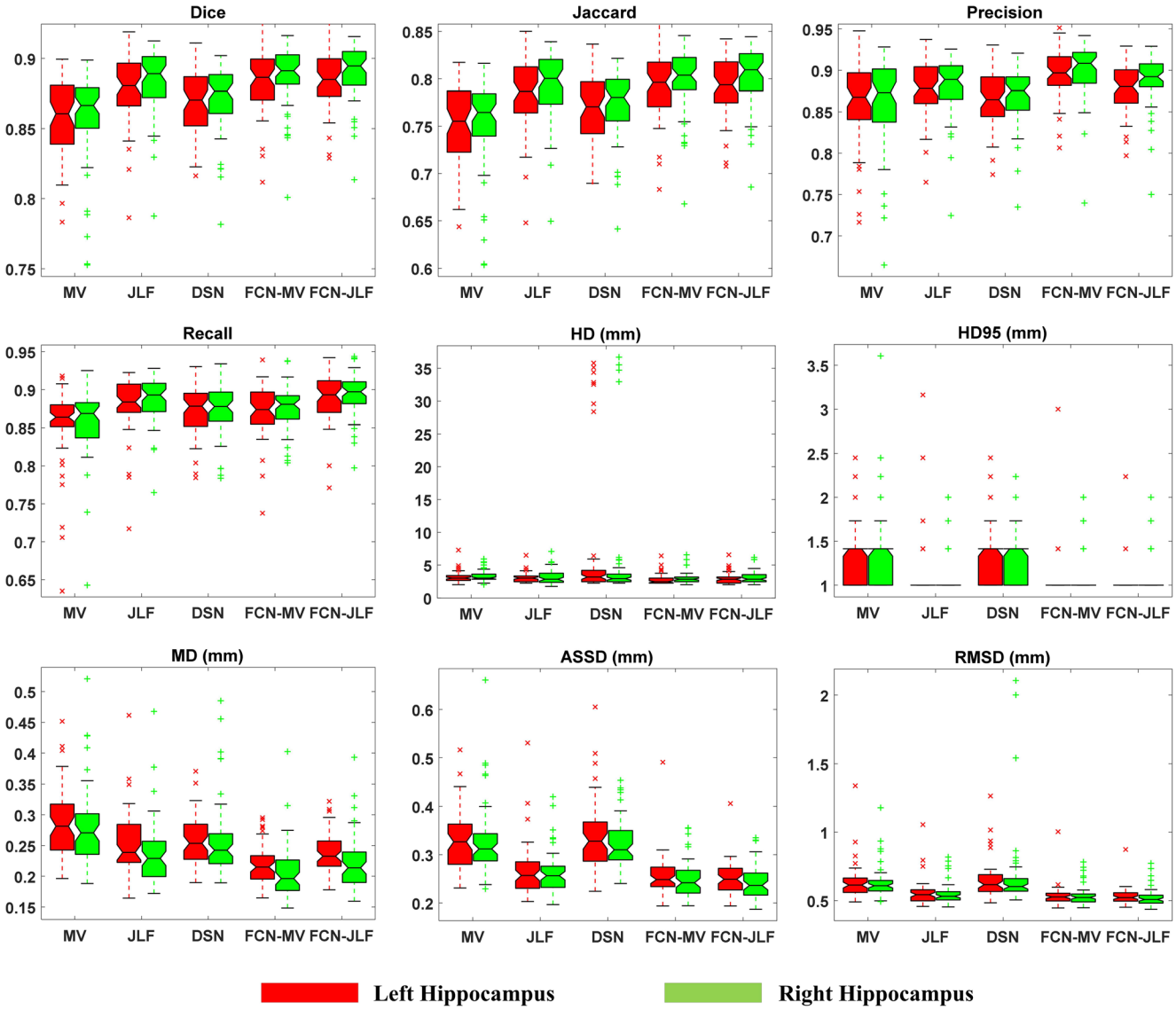

Fig. 4 Box plots of the segmentation results based on nine evaluation measures. In each box, the central mark is the median, and edges are the 25th and 75th percentiles, respectively

For DSN, we select 40 subjects (i.e., the subjects used as atlases in the proposed methods) as training set and other 60 subjects as testing set. During training, 4 subjects were randomly selected for validation. As the restriction of GPU memory, we use image patches as input for the network, instead of using the whole images. The patch size is set to $16 \times 16 \times 16$, optimally selected from $8 \times 8 \times 8$, $16 \times 16 \times 16$, $24 \times 24 \times 24$. We separately construct the training set and then train separate DSN models for the left and right hippocampi, respectively.

Table 2 lists the nine index values of segmentation results using different segmentation methods. It shows that our proposed methods, FCN-MV and FCN-JLF, obtain the best results. Compared with the MV method, FCN-MV improves the Dice scores by 2.7% and 2.8% for the left and right hippocampus segmentation results. This improvement is achieved by the FCN based confidence estimation, which potentially compensates for the registration error. FCN-JLF can further improve the Dice scores by 0.1% and 0.3% for the left and right hippocampus segmentation results, compared with FCN-MV. This improvement is achieved by using the more advanced label fusion method, JLF, for fusing the corrected label maps. It can also be observed that JLF improves MV by 2.4% both for the left and right hippocampi, while FCN-JLF only improves FCN-MV 0.1% and 0.3% for the left and right hippocampi. This demonstrates that our proposed FCN based label correction method can effectively correct registration errors. With label correction, even the simplest majority voting label fusion can achieve better segmentation results than the state-of-the-art JLF method. Our proposed methods also obtain better segmentation results than the deep learning segmentation method (DSN).

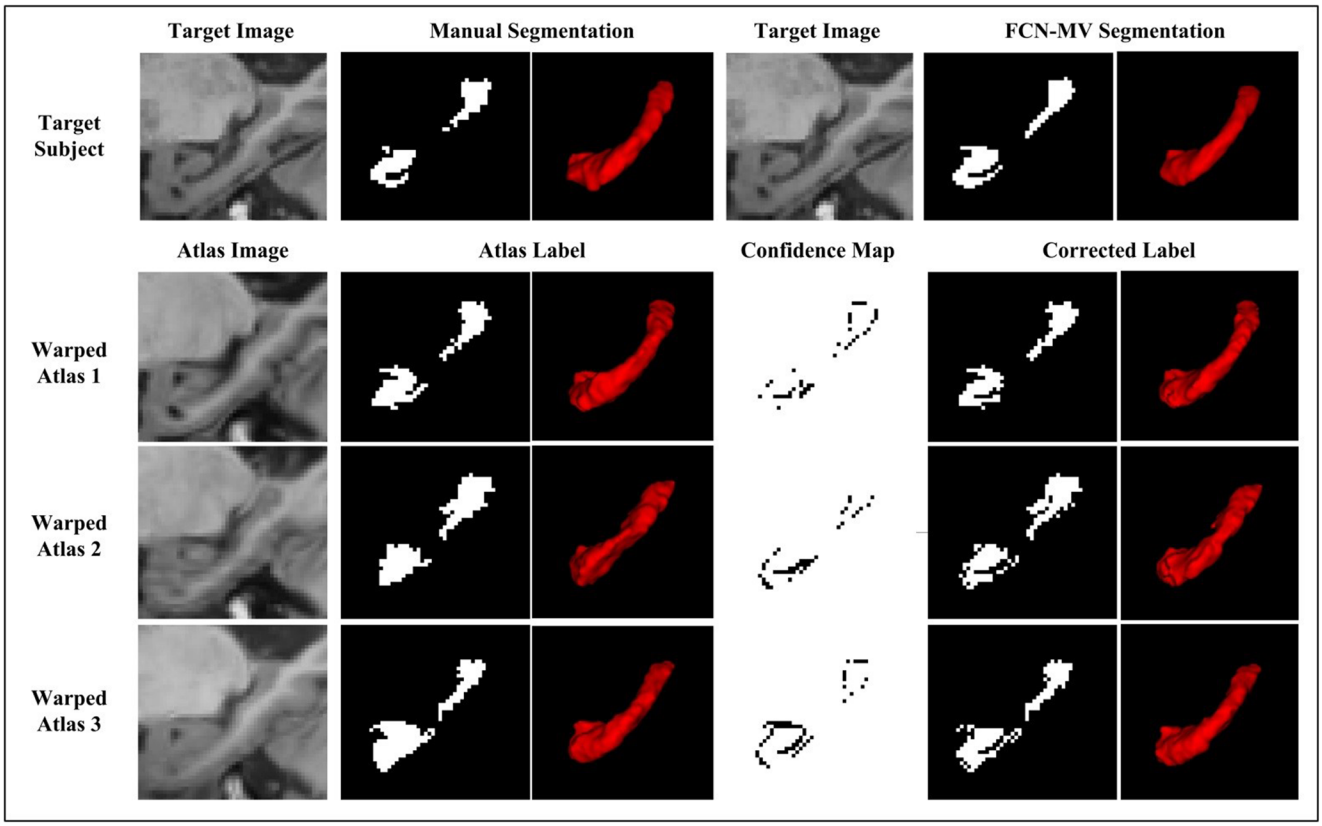

Fig. 5 Examples of confidence maps and corrected (warped) atlas label maps

Figure 4 shows the box plots of the segmentation results based on the nine evaluation measures. It is obvious that our proposed methods perform consistently better than other methods. For the HD measure, several severe outliers can be observed in the segmentation results obtained by DSN. For the HD95 measure, one can see in the figure that the boxes turn to lines for both left and right hippocampus segmentation results obtained by our proposed methods. This means that the HD95 values at the 25th and 75th percentiles reach the same value, indicating that our proposed method is very robust. The similar results are obtained by JLF method.

Figure 5 shows examples of confidence maps and corrected (warped) atlas label maps. In the confidence maps, dark voxels denote the confidence values of 0, which means that registration errors may happen at these voxels. The corrected atlas label maps are obtained by changing the label values at the voxels with the confidence values being 0. We can see that the corrected (warped) atlas label maps are more similar to the target label, compared to the original warped atlas labels. Meanwhile, we can also find some artifacts in the second corrected atlas label, which makes the label unsmooth. Interestingly, most of these artifacts can be perfectly

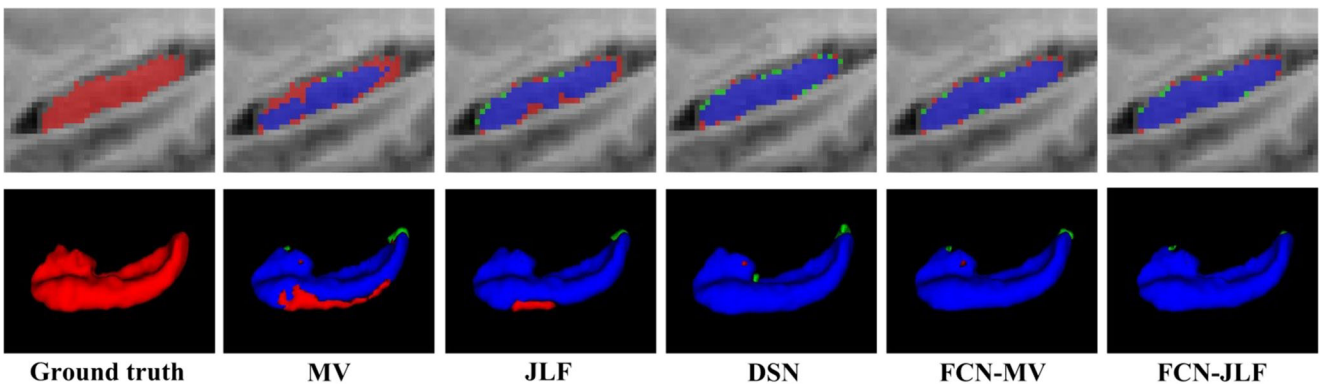

Fig. 6 Sagittal view (top row) and 3D rendering (bottom row) of left hippocampus segmentations for a randomly selected subject (red: ground-truth; green: automated segmentations; blue: overlap between ground-truth and automated segmentations)

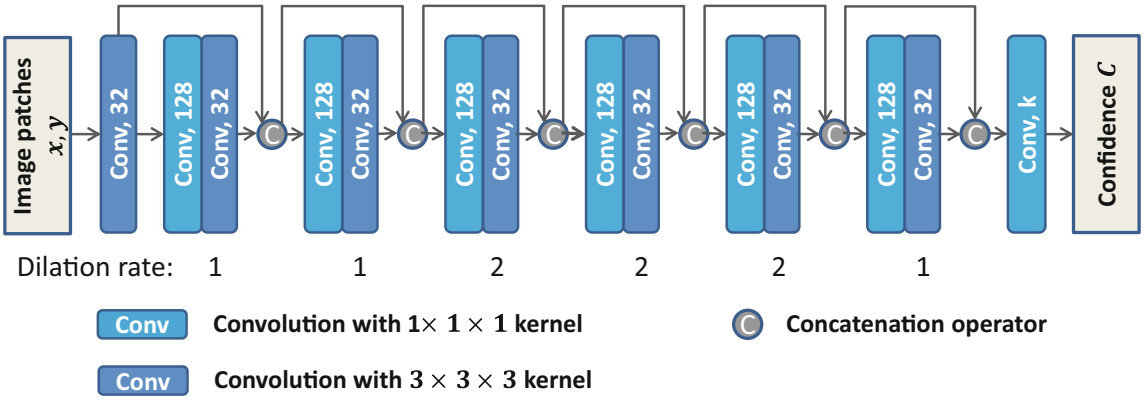

Fig. 7 Illustration of the dilated dense network structure. The number of kernels is denoted in each convolution operation rectangle

eliminated by label fusion, resulting smooth segmentation, which can be observed in the FCN-MV segmentation from the figure. Figure 6 shows sagittal view and 3D rendering of left hippocampus segmentations for a randomly selected subject. We can observe that our proposed methods produce the most accurate segmentation results.

Discussion

Multi-atlas image segmentation (MAIS) has recently gained lots of attention for medical image segmentation (Iglesias and Sabuncu 2015), in which a set of atlases are first selected and then registered to the target image. Next, the corresponding atlas labels are warped to the target image space and further combined to obtain the segmentation (i.e., the label fusion step). This step is very important in the MAIS method, as it deals with registration errors that may reduce accuracy and introduce unnecessary smoothness in the segmentation results. One of the most popular label fusion methods is the local weighted voting method. The pre-defined similarity functions are often directly used for estimating the confidence in the local weighted voting method, such as Gaussian function (Sabuncu et al. 2010) and inverse function (Artaechevarria et al. 2009).

The drawback of using a pre-defined similarity function to estimate the confidence of each atlas is sub-optimality of the obtained confidence, as two similar (*in terms of* appearance) patches may belong to different tissue classes (Bai et al. 2015). To overcome this drawback, similar to those local supervised models for learning the confidence of two patches with the same class label (Benkarim et al. 2017; Zhu et al. 2017), we have developed a deep learning based method by using a fully convolutional network (FCN) to robustly estimate the confidence values. Unlike previous supervised methods (e.g., (Benkarim et al. 2017; Zhu et al. 2017)) that require separate training for each voxel, which is computationally too complex, our method trains a single model for all voxel locations. With the estimated confidence, even a simple majority voting method can obtain superior segmentation results, compared to all other widely-used state-of-the-art methods.

According to (Zhu et al. 2017), patch-wise label fusion strategies often obtain more accurate segmentation results compared to the voxel-wise label fusion methods. Thus, we implemented our proposed method as a patch-wise label fusion strategy by estimating the confidence map of the whole image patch, instead of only estimating confidence for the center voxel. This was implemented using a new architecture of fully convolutional networks (Long et al. 2015). Specifically, we used the residual U-Net as the base architecture (Milletari et al. 2016; Chen et al. 2018). U-Net consists of

Table 3 Nine index values (mean ±std) of hippocampus segmentation results using different methods

	FCN-MV	DDN-MV
Dice (L/R)	0.883 ± 0.022 / 0.888 ± 0.021	0.884 ± 0.022 / 0.889 ± 0.021
Jaccard (L/R)	0.792 ± 0.034 / 0.800 ± 0.033	0.792 ± 0.034 / 0.800 ± 0.032
Precision (L/R)	0.896 ± 0.029 / 0.902 ± 0.033	0.897 ± 0.029 / 0.901 ± 0.031
Recall (L/R)	0.872 ± 0.033 / 0.876 ± 0.028	0.872 ± 0.034 / 0.878 ± 0.028
HD (L/R)	2.843 ± 0.770 / 3.013 ± 0.878	2.836 ± 0.719 / 3.013 ± 0.904
HD95 (L/R)	1.054 ± 0.271 / 1.099 ± 0.244	1.064 ± 0.295 / 1.097 ± 0.217
MD (L/R)	0.218 ± 0.032 / 0.205 ± 0.044	0.216 ± 0.032 / 0.206 ± 0.042
ASSD (L/R)	0.255 ± 0.042 / 0.249 ± 0.036	0.254 ± 0.044 / 0.249 ± 0.037
RMSD (L/R)	0.533 ± 0.073 / 0.534 ± 0.067	0.533 ± 0.076 / 0.532 ± 0.067

a down-sampling path and an up-sampling path. The down-sampling path alternates with convolution layers and pooling layers to enlarge the receptive field. The up-sampling path alternates with convolution layers and deconvolution layers to recover the image resolution. The feature maps in the down-sampling path are concatenated to the corresponding feature maps in the up-sampling path with long-skip connections, aggregating multi-scale features for dense prediction. We also used residual connections to group every two convolution layers, which can promote the information flow and accelerate the convergence (He et al. 2016b).

Besides the U-Net structures, dilated convolution networks can also be used to enlarge the receptive field (Yu and Koltun 2015), and dense connection networks can be used to aggregate multi-scale features (Huang et al. 2017). Combining these two network structures, the dilated dense networks have been recently proposed and successfully applied in different applications (Xu et al. 2019; Shamsolmoali et al. 2019). To illustrate the effectiveness of our residual U-Net structure, we replaced it with a dilated dense network (DDN) in our proposed framework for label correction. The structure of the dilated dense network is shown in Fig. 7. Table 3 shows nine index values of the segmentation results obtained by FCN-MV and DDN-MV. It can be observed that the segmentation results obtained by these two methods are very similar. This demonstrates that our residual U-Net structure is as effective as the recently proposed dilated dense network structure for label correction.

We also compared the proposed methods to a state-of-the-art deep learning based image segmentation method with 3D deeply supervised network (DSN) (Dou et al. 2017). Because of the restriction of limited GPU memory, DSN used image patches as input, ignoring the global spatial information. This led to some artifacts in the segmentation results, such as isolated points and holes. Figure 8 shows two hippocampus segmentations obtained by DSN, in which such artifacts can be noticed.

In our proposed methods, the inputs to our architecture were an image patch from the target image and the corresponding patch from each warped atlas. With this formulation, our confidence estimation can be relatively easily completed, and does not require as much global image information as the semantic segmentation methods. In fact, image registration between each atlas image and the target image has already captured the global spatial information of brain structure, and the label fusion step can further eliminate possible artifacts introduced by label correction to make the segmentation smooth.

As shown in Table 2 and Fig. 4, our proposed methods obtain the smallest standard deviations in most segmentation evaluation measures. For the segmentation results of DSN, because of the artifacts, several severe outliers can be observed in HD measure, which demonstrates that 3D deep learning segmentation method DSN is not as robust as our proposed methods. Thus, by combining the advantages of deep learning methods and multi-atlas segmentation methods, our proposed methods can obtain more accurate and robust segmentation results.

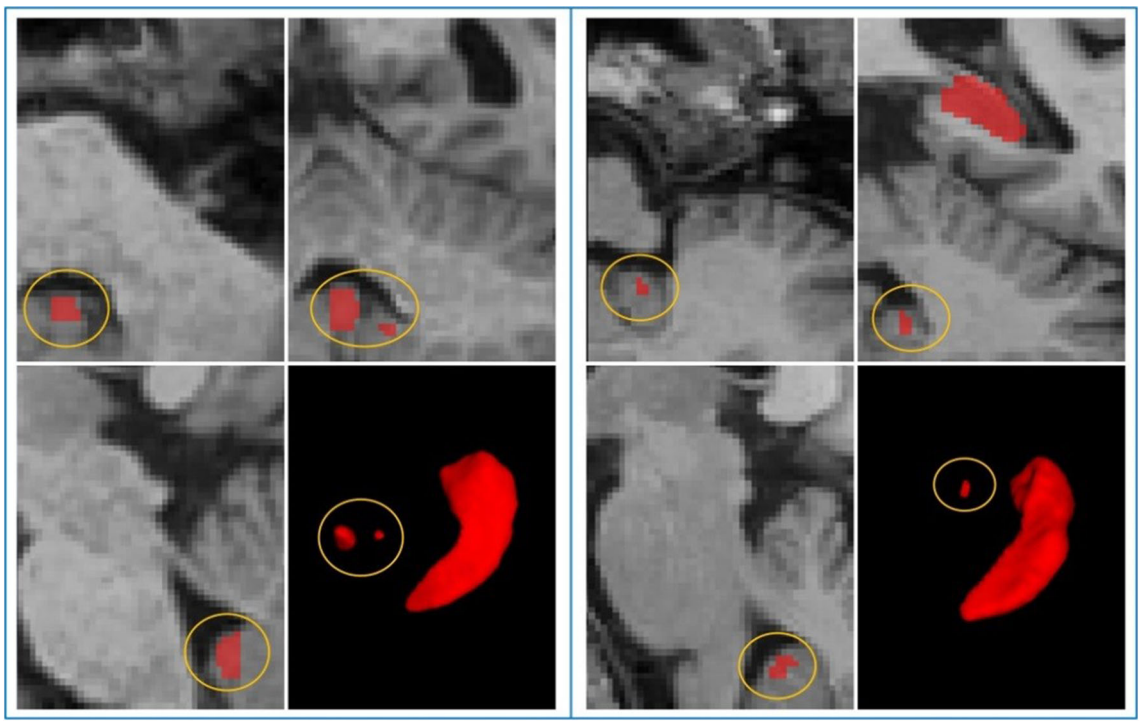

Fig. 8 Two selected hippocampus segmentation results obtained by DSN

Conclusion

In this paper, we have presented a new multi-atlas label fusion framework by using deep learning for confidence estimation. Specifically, deep learning was used to identify the potential errors in the warped atlas labels and they were then corrected based on the estimated confidence maps. The final segmentation was obtained by two label fusion methods, MV and JLF, on those corrected (warped) atlas labels. Our proposed methods, FCN-MV and FCN-JLF, have been validated on a public dataset for hippocampus segmentation. The results show better performance of our proposed methods than the state-of-the-art segmentation methods.

Information Sharing Statement

MR images used in this manuscript can be freely downloaded from Alzheimer's Disease Neuroimaging Initiative (ADNI) database (RRID:RRID:SCR_003007, http://adni.loni.usc.edu/). Ground-truth hippocampus labels of the image data were provided by the European Alzheimer's Disease Consortium and Alzheimer's Disease Neuroimaging Initiative, and can be freely downloaded (www.hippocampal-protocol.net). Software developed in this manuscript is available upon request from Dr. Zhu (Email: hancanzhu@yeah.net).

Acknowledgments This work was supported in part by National Natural Science Foundation of China [61602307, 61877039] and Natural Science Foundation of Zhejiang Province [LY19F020013, LY20F020011].

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J. V., & Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage, 46*(3), 726–738.

Artaechevarria, X., Muñoz-Barrutia, A., and Ortiz-de-Solorzano, C.(2008). "Efficient classifier generation and weighted voting for atlas-based segmentation: Two small steps faster and closer to the combination oracle," *SPIE Medical Imaging*, 69141W–69141W-9.

Artaechevarria, X., Munoz-Barrutia, A., & Ortiz-de-Solorzano, C. (2009). Combination strategies in multi-atlas image segmentation: Application to brain MR data. *Medical Imaging, IEEE Transactions on, 28*(8), 1266–1277.

Asman, A. J., & Landman, B. A. (2012). Formulating spatially varying performance in the statistical fusion framework. *Medical Imaging, IEEE Transactions on, 31*(6), 1326–1336.

Asman, A. J., & Landman, B. A. (2013). Non-local statistical label fusion for multi-atlas segmentation. *Medical Image Analysis, 17*(2), 194–208.

Asman, A. J., & Landman, B. A. (2014). Hierarchical performance estimation in the statistical label fusion framework. *Medical Image Analysis, 18*(7), 1070–1081.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis, 12*(1), 26–41.

Bai, W., Shi, W., O'Regan, D. P., et al. (2013). A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images. *Medical Imaging, IEEE Transactions on, 32*(7), 1302–1315.

Bai, W., Shi, W., Ledig, C., & Rueckert, D. (2015). Multi-atlas segmentation with augmented features for cardiac MR images. *Medical Image Analysis, 19*(1), 98–109.

Benkarim, O. M., Piella, G., Ballester, M. A. G., et al. (2017). Discriminative confidence estimation for probabilistic multi-atlas label fusion. *Medical Image Analysis, 42*, 274–287.

Boccardi, M., Bocchetta, M., Morency, F. C., Collins, D. L., Nishikawa, M., Ganzola, R., Grothe, M. J., Wolf, D., Redolfi, A., Pievani, M., Antelmi, L., Fellgiebel, A., Matsuda, H., Teipel, S., Duchesne, S., Jack CR Jr, Frisoni, G. B., & EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer's Disease Neuroimaging Initiative. (2015). Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimers Dement, 11*(2), 175–183.

Cao, Y., Yuan, Y., Li, X. et al. (2011). "Segmenting images by combining selected atlases on manifold," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 272–279.

Chen, H., Dou, Q., Yu, L., Qin, J., & Heng, P. A. (2018). VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage, 170*, 446–455.

Commowick, O., Akhondi-Asl, A., & Warfield, S. K. (2012). Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE. *Medical Imaging, IEEE Transactions on, 31*(8), 1593–1606.

Coupé, P., Manjón, J. V., Fonov, V., Pruessner, J., Robles, M., & Collins, D. L. (2011). Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage, 54*(2), 940–954.

Doshi, J., Erus, G., Ou, Y., Resnick, S. M., Gur, R. C., Gur, R. E., Satterthwaite, T. D., Furth, S., Davatzikos, C., & Alzheimer's Neuroimaging Initiative. (2016). MUSE: MUlti-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage, 127*, 186–195.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., & Heng, P. A. (2017). 3D deeply supervised network for automated segmentation of volumetric medical images. *Medical Image Analysis, 41*, 40–54.

A. K. H. Duc, M. Modat, K. K. Leung et al., "Manifold learning for atlas selection in multi-atlas-based segmentation of hippocampus," Medical Imaging 2012: Image Processing, 8314, 83140Z (2012).

Fang, L., Zhang, L., Nie, D. et al. (2017). "Brain Image Labeling Using Multi-atlas Guided 3D Fully Convolutional Networks," *International Workshop on Patch-based Techniques in Medical Imaging*, 12–19.

Gu, J., Wang, Z., Kuen, J., et al. (2017). Recent advances in convolutional neural networks. *Pattern Recognition, 77*, 354–377.

Haber, E., & Modersitzki, J. (2004). Numerical methods for volume preserving image registration. *Inverse Problems, 20*(5), 1621.

Hao, Y., Wang, T., Zhang, X., Duan, Y., Yu, C., Jiang, T., Fan, Y., & Alzheimer's Disease Neuroimaging Initiative. (2014). Local label learning (LLL) for subcortical structure segmentation: Application to hippocampus segmentation. *Human Brain Mapping, 35*(6), 2674–2697.

Haom, Y., Liu, J., Duan, Y. et al. (2012). "Local label learning (L3) for multi-atlas based segmentation," *SPIE Medical Imaging*, 83142E-83142E-8.

He, K., Zhang, X., Ren, S. et al. (2016a). "Identity mappings in deep residual networks," *European Conference on Computer Vision*, 630–645.

He, K., Zhang, X., Ren, S. et al. (2016b). "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., & Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage, 33*(1), 115–126.

Huang, G., Liu, Z., Van Der Maaten, L. et al. (2017). "Densely connected convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Iglesias, J. E., & Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis, 24*(1), 205–219.

Jack, C. R., Bernstein, M. A., Fox, N. C., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging, 27*(4), 685–691.

Jafari-Khouzani, K., Elisevich, K. V., Patel, S., & Soltanian-Zadeh, H. (2011). Dataset of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques. *Neuroinformatics, 9*(4), 335–346.

Jia, Y., Shelhamer, E., Donahue, J. et al. (2014). "Caffe: Convolutional architecture for fast feature embedding," *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678.

Jorge Cardoso, M., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N. C., Ourselin, S., & Alzheimer's Disease Neuroimaging Initiative. (2013). STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Medical Image Analysis, 17*(6), 671–684.

Langerak, T. R., Berendsen, F. F., Van der Heide, U. A., et al. (2013). Multiatlas-based segmentation with preregistration atlas selection. *Medical Physics, 40*(9), 091701.

Liao, S., Gao, Y., & Shen, D. (2012). Sparse patch based prostate segmentation in CT images. *Medical Image Computing and Computer-Assisted Intervention–MICCAI, 2012*, 385–392.

Liao, S., Gao, Y., Lian, J., et al. (2013). Sparse patch-based label propagation for accurate prostate localization in CT images. *Medical Imaging, IEEE Transactions on, 32*(2), 419–434.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lötjönen, J. M. P., Wolz, R., Koikkalainen, J. R., et al. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage, 49*(3), 2352–2365.

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *3D Vision (3DV), 2016 Fourth International Conference on*, 565–571.

Rohlfing, T., Brandt, R., Menzel, R., & Maurer CR Jr. (2004). Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage, 21*(4), 1428–1442.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Rousseau, F., Habas, P. A., & Studholme, C. (2011). A supervised patch-based approach for human brain labeling. *Medical Imaging, IEEE Transactions on, 30*(10), 1852–1862.

Sabuncu, M. R., Yeo, B. T. T., Van Leemput, K., et al. (2010). A generative model for image segmentation based on label fusion. *Medical Imaging, IEEE Transactions on, 29*(10), 1714–1729.

Sanroma, G., Wu, G., Gao, Y., et al. (2014). Learning to rank atlases for multiple-atlas segmentation. *Medical Imaging, IEEE Transactions on, 33*(10), 1939–1953.

Shamsolmoali, P., Zhang, J., & Yang, J. (2019). Image super resolution by dilated dense progressive network. *Image and Vision Computing, 88*, 9–18.

Wang, H., Suh, J. W., Das, S. et al. (2011). "Regression-based label fusion for multi-atlas segmentation," *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1113–1120.

Wang, H., Suh, J. W., Das, S. R., et al. (2013). Multi-atlas segmentation with joint label fusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35*(3), 611–623.

Warfield, S. K., Zou, K. H., & Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *Medical Imaging, IEEE Transactions on, 23*(7), 903–921.

Xu, B., Ye, H., Zheng, Y., Wang, H., Luwang, T., & Jiang, Y. G. (2019). Dense dilated network for video action recognition. *IEEE Transactions on Image Processing, 28*(10), 4941–4953.

Yang, H., Sun, J., Li, H. et al. (2017). "Neural Multi-Atlas Label Fusion: Application to Cardiac MR Images," arXiv preprint arXiv: 1709.09641.

Yu, F., and Koltun, V. (2015). "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122.

Yu, L., Yang, X., Chen, H. et al. (2017). "Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images," *AAAI*, 66–72.

Zaffino, P., Ciardo, D., Raudaschl, P., et al. (2018). Multi atlas based segmentation: Should we prefer the best atlas group over the group of best atlases? *Physics in Medicine & Biology, 63*(12), 12NT01.

Zhu, H., Cheng, H., and Fan, Y. (2015). "Random local binary pattern based label learning for multi-atlas segmentation," *SPIE Medical Imaging*, 94131B-94131B-8.

Zhu, H., Cheng, H., Yang, X., Fan, Y., & Alzheimer's Disease Neuroimaging Initiative. (2017). Metric learning for multi-atlas based segmentation of Hippocampus. *Neuroinformatics, 15*(1), 41–50.