Automatic Brain Labeling via Multi-Atlas Guided Fully Convolutional Networks

Fang, Longwei; Zhang, Lichi; Nie, Dong; Cao, Xiaohuan; Rekik, Islem; Lee, Seong-Whan

Published in:
Medical Image Analysis

DOI:
10.1016/j.media.2018.10.012

Publication date:
2019

Automatic Brain Labeling via Multi-Atlas Guided Fully Convolutional Networks

Longwei Fang[a,b,e], Lichi Zhang[d,e], Dong Nie[e], Xiaohuan Cao[e,g], Islem Rekik[h], Seong-Whan Lee[f], Huiguang He[a,b,c,*], Dinggang Shen[e,f,*]

[a] Research Center for Brain-inspired Intelligence and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences(CAS), Beijing, 100190, China

[b] *University of Chinese Academy of Sciences, Beijing, China*

[c] *Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China*

[d] *Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China*

[e] *Department of Radiology and BRIC, University of North Carolina at Chapel Hill, North Carolina, USA*

[f] *Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea*

[g] *School of Automation, Northwestern Polytechnical University, Xi'an, China*

[h] *BASIRA lab, CVIP, School of Science and Engineering, Computing, University of Dundee, UK*

{ dgshen@med.unc.edu, huiguang.he@ia.ac.cn}

** Corresponding authors*

Abstract

Multi-atlas-based methods are commonly used for MR brain image labeling, which alleviates the burdening and time-consuming task of manual labeling in neuroimaging analysis studies. Traditionally, multi-atlas-based methods first register multiple atlases to the target image, and then propagate the labels from the labeled atlases to the unlabeled target image. However, the registration step involves non-rigid alignment, which is often time-consuming and might lack high accuracy. Alternatively, patch-based methods have shown promise in relaxing the demand for accurate registration, but they often require the use of hand-crafted features. Recently, deep learning techniques have demonstrated their effectiveness in image labeling, by automatically learning comprehensive appearance features from training images. In this paper, we propose a *multi-atlas guided fully convolutional network (MA-FCN)* for automatic image labeling, which aims at further improving the labeling performance with the aid of prior knowledge from the training atlases. Specifically, we train our MA-FCN model in a patch-based manner, where the input data consists of *not only* a training image patch *but also* a set of its neighboring (i.e., most similar) affine-aligned atlas patches. The guidance information from neighboring atlas patches can help boost the discriminative ability of the learned FCN. Experimental results on different datasets demonstrate the effectiveness of our proposed method, by significantly outperforming the conventional FCN and several state-of-the-art MR brain labeling methods.

Keywords: Brain image labeling, multi-atlas-based method, fully convolutional network, patch-based labeling

1.Introduction

Anatomical brain labeling is highly desired for region-based analysis of MR brain images, which is important for many research studies and clinical applications, such as facilitating diagnosis [1, 2] and investigating early brain development [3]. Also, brain labeling is a fundamental step in brain network analysis pipelines, where regions-of-interest (ROIs) need to be identified prior to exploring any connectivity traits [4-7]. But it is labor-intensive and impractical to manually label a large set of 3D MR images, thus recent developments focused on automatic labeling of brain anatomy. However, there are multiple challenges in automatic labeling: 1) complex brain structures, 2) ambiguous boundaries between neighboring regions as observed by the highlighted region in Figure 1, and 3) large variation of the same brain structure across different subjects.

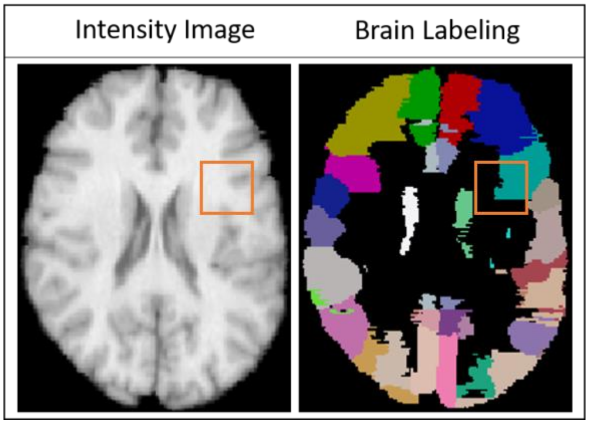

Figure 1: Typical example of brain MR intensity image (left) and its label map (right). The region inside the orange rectangle has a blurry boundary, which is challenging for automatic brain labeling.

Recently, many attempts have been made to address these challenges in MR brain labeling [8-15]. In particular, the multi-atlas-based labeling methods have been widely used as standard approaches for their effectiveness and robustness. Basically, through defining an atlas as a combination of the intensity image with its manually-labeled map, one can label a target image in two steps: 1) registering the atlas image to the target image, and then 2) propagating the atlas label map to the target image. This generalizes to multi-atlas labeling methods, where multiples atlases are first registered to the target image, and then labels from all labeled atlases are propagated to the target unlabeled image. Generally, the multi-atlas-based methods can be classified into two categories: *registration-based* and *patch-based* methods. Typically, *registration-based* methods first align multiple atlases to the target image in the registration step [16, 17], and then fuse the respective warped atlas label maps to obtain the final labels in the label fusion step [8, 18-20]. The main drawback of such methods is that the labeling performance highly depends on the reliability of non-rigid registration techniques used, which is often quite time-consuming [21].

Patch-based methods, on the other hand, have gained increased attention in image labeling, since they can alleviate the need for high registration accuracy through exploring several neighboring patches within a local search region [22-27]. For such methods, affine registration of the atlases to the target image is often used. Specifically, for each target patch, similar patches are selected from the affine-aligned atlas images according to patch similarities within a search region. Then, the labels of those selected atlas patches are fused together to label the subject patch. The underlying assumption of patch-based methods is that, when two patches are similar in intensity, they are also similar in labels [28]. To measure the similarity between patches, several feature extraction methods have been proposed based on anatomical structures [22, 29] or intensity distributions [23, 24]. However, these hand-crafted patch-driven features have a key limitation. For example, they are limited by using a pre-defined set of features (i.e., color, gradient, shape, intensity distribution etc.), without exploring other possible features that can be considered and learned when comparing patches for our target task.

Recently, the convolutional networks (ConvNet) methods have shown great promise and performance in several medical image analysis tasks, including image segmentation [30-33] and image synthesis [34-36]. An appealing aspect of ConvNet is that it can automatically learn the most comprehensive, high-level appearance features that can best represent the image.

Specifically, the fully convolutional network (FCN) [37] have demonstrated its effectiveness in medical image segmentation. For example, Nie *et al.* [38] adopted the FCN model for brain tissue segmentation, which significantly outperformed the conventional segmentation methods in terms of accuracy.

In this paper, we propose a novel *multi-atlas guided fully convolution network (MA-FCN)* aiming at further improving the labeling performance with the aid of patch-based manner and the registration-based labeling. To guide the learning of a conventional FCN for automatic brain labeling by leveraging available multiple atlases, we align a subset of the training atlases to the target images. Note that we only implement affine registration (with 12 degree of freedom using normalized correlation as cost function) to roughly align atlases to the target image, instead of non-rigid registration, which ensures efficiency and also demonstrates the ability of the FCN for inferring labels from local regions. In the training stage, we propose a novel candidate target patch selection strategy for helping identify the optimal set of candidate target patches, thus balancing the large variability of ROI sizes. Both target patches and their corresponding candidate atlas patches (two training sources) are used for training the FCN model. We take our proposed FCN model one step further by devising three novel strategies to incorporate the extracted appearance features from the two training sources in a more effective way, i.e., atlas-unique pathway, target-patch pathway, and atlas-aware fusion pathway. Specifically, atlas-unique pathway and target-patch pathway process the atlas patch and target patch separately, while atlas-aware fusion pathway merges these pathways together. The main contributions of our method are two-fold:

(1) We guide the learning of FCN model by leveraging the available information in multiple atlases.

(2) The proposed method does not need a non-rigid registration step for aligning atlases to the target image, which is more efficient for brain labeling.

2. Related Works

Registration-based labeling. Registration based methods leverage both non-linear registration and label fusion techniques. Many relevant works were proposed to improve the performance of the registration step, including the LEAP method [39] which constructs an image manifold according to the similarities between all training and test images. The sophisticated tree-based group-wise registration strategy developed in [40] employed pairwise registration strategy that concatenated precomputed registrations between pairs of atlases (Wang et al. 2013). For the label fusion step, the voting-based strategies proposed by [8, 41-47] are popular for fusing the warped atlas labels. For instance, Langerak *et al.* [8] defined a global weight for each atlas by its similarity in intensity to the target image, and then performed a weighted sum of all atlas labels to get the final label. They used a single weight for the whole atlas image, which overlooks the fact that subject-to-subject similarity varies across anatomical regions. To address this limitation, Artaechevarria *et al.* [42] proposed a local weighted voting method to fuse weights in a voxel-wise manner. Specifically, the weight of each voxel is computed using the mutual information similarity of the atlas image and the target image in a small region. The local weighted strategy can boost the accuracy of label propagation; however, it may fail in highly variable anatomical regions that cannot be simultaneously captured by *all* atlases. To avoid this limitation, Isgum *et al.* [43] used an atlas selection strategy to select a subset of atlases with the highest similarities to the target image by statistical pattern recognition theory. Then, the propagated labels were combined by spatially varying decision fusion weights. In a different work, Sanroma et al. [48] combined a learning-based atlas selection strategy with nonlocal weighted voting to label a brain. The best atlases were selected based on their expected labeling accuracy by learning the relationship between the pairwise appearance of the observed instances and their final labeling performance, and then the final label value was voted from both local and neighboring voxels in the selected atlases. The limitation of this method is that the weights are computed independently for each atlas, without taking into account the fact that different atlases may produce similar label errors. Wang et al. [20] solved this limitation by proposing a joint label fusion strategy (JLF), in which joint probability of pairwise atlases is modeled to estimate the segmentation error at a voxel, and then weighted voting is formulated in terms of minimizing the total expectation of labeling error. One major limitation of registration-based methods is that it takes lots of time to align atlases to the target image.

Patch-based labeling. Patch-based labeling methods use a non-local strategy to alleviate the need for high registration accuracy. They propagate the label information of the selected similar atlas patches, which are identified within a local neighborhood of the target patch. Most patch based methods are constructed assuming only affine registration as a prerequisite to align the atlases to the target image because affine registration is much faster than non-rigid registration. Some methods use sparse patch selection strategy to select the most similar intensity patches for the target training patch to improve the label fusion step. Zhang et al. [49] segmented the brain by using a sparse patch-based label fusion (SPBL) strategy. Candidate image patches are selected from a neighborhood region to build a graph, and then a sparse constraint is applied to the candidate atlas patches to derive the graph weights. Finally, the patches are fused together by a weighted fusion function. In other works, the learning strategies are proposed to learn the mapping from the input intensity patch to the final label map. Zhang et al. [29] proposed to label the brain by using a hierarchical random forest. They clustered similar patches together to learn a bottom-level forest, and then the bottom-level forests were clustered together by their capabilities. Finally, the high-level forest was trained by clustering bottom-level forests and all atlases. The limitation of their method is that the performance can be easily influenced by the cluster strategy. Zikic et al. [24] proposed to build atlas forests (AF) by using a small and deep classification forest, which encodes each atlas individually in reference to an aligned probabilistic atlas map. Each atlas forest produces one probability label estimation, and then all label estimations are averaged to get the final label. Their method is fast since only one registration is needed to align the target image to the probabilistic atlas map. However, this method requires manually designed features to train the forest, without exploring other possible image features, which may not best represent the target image. Some methods combine registration-based method with patch-based method together to improve the labeling performance. Wu et al. [11] proposed a hierarchical feature representation and label-specific patch partition method (HSPBL), which is a combination of registration-based method and patch-based method. Specifically, they use non-rigid registration to preprocess the atlas data, and then each image patch is represented by multi-scale features that encode both local and semi-local image information to increase the fidelity of similarity calculation. Finally, the atlas patch is further partitioned into a set of label-specific partial image patches by atlas label information.

ConvNet labeling. ConvNet, on the other hand, can automatically learn the high-level features of the image. One of the widely used ConvNet architectures in image labeling is convolutional neural networks (CNN) [50, 51], which learns convolution kernels to simulate the receptive fields of our visual system [52] and extracts the deep features from the image. The parameters of the convolution kernels are updated by back-propagation of the errors. However, CNN is limited by a lack of efficiency in processing the whole brain image as it uses a patch-to-voxel prediction strategy, which can only predict the label of a center voxel for each input patch. To solve this issue, fully convolutional networks (FCN) [37, 38] were developed by using a patch-to-patch training strategy without using the fully connected layer. FCN typically inputs a patch and outputs the predicted label of the whole patch. U-Net [30] and V-Net [31] were also introduced to label brains by combining shallow layers with corresponding deep layers in FCN. This allows merging learned features at different depths of the network and helps avoid gradient degeneration when reaching shallow layers, thus guaranteeing the convergence of the network training.

3. Method

In this section, we detail the proposed MA-FCN framework for automatic brain labeling. Our goal is to improve the labeling performance of a typical FCN by guiding and boosting its learning using multiple aligned atlases. Our method comprises *training* and *testing* stages. In the *training* stage, we randomly select several training images as atlases. Specifically, we first select 3D patches from the training images using a random selection strategy. Next, for each selected training 3D patch, we select the K most similar candidate atlas patches within a specific search window. Then, all training patches and their corresponding selected candidate atlas patches are input into the MA-FCN model for training. Note that the atlas patch refers to the combination of atlas intensity patch and its corresponding label patch. In the *testing* stage, each testing 3D patch is concatenated with its K most similar atlas patches, and then fed into MA-FCN to predict the label patch. Since each target voxel $\boldsymbol{x}$ in the brain belongs to many overlapping 3D patches, we fuse all the predicted labels from all patches containing $\boldsymbol{x}$ to finally label the target voxel by majority voting.

3.1. Data Preparation

Prior to the atlas patch selection step, we affine register all atlases (i.e., intensity images and their corresponding label maps) to the training data using FLIRT in FSL toolkit [53]. Next, we propose a patch sampling and selection strategy to identify the most similar atlas patch to the target patch. Figure 2 presents the flowchart of our novel strategies for training patch sampling and atlas patch selection, which are further detailed in Sections 3.1.1 and 3.1.2, respectively.

3.1.1. Training patch sampling

Noting the large variability in size across anatomical ROIs, randomly sampling from the whole brain will create an imbalance in training samples across different ROIs. For instance, a whole-brain sampling strategy might select many more locations within large ROIs than smaller ones, which will weaken the model learning for small brain anatomical regions. On the other hand, ROI boundaries are very important in labeling since they contain direct structural information, but voxels near the boundaries are more difficult to classify than the inside voxels. Therefore, more training samples should be sampled along the boundaries of the target ROIs.

We proposed a boundary-focused patch extraction strategy to solve the imbalance samples by randomly sampling patches across the whole brain. For each labeled ROI, we detect its boundary using the Canny edge detector, thereby creating an edge map for each target intensity image (Figure 2). We also extract the inner voxels within each ROI while excluding the edge to build an inner voxel location map. Then, we randomly sample locations from both edge and inner voxel maps while ensuring that: 1) the number of samples extracted from each ROI is the same, and 2) the number of patches extracted around the boundary is larger than that from the inside of each ROI. In our experiment, the ratio between the boundary and inside patches is set to 4:1. We have tested the ratios 1:1 and 2:1 and found that the performance of 2:1 is better than 1:1. Then we tested the ratio 4:1 and found that it has the same performance as 2:1. Thus, we choose ratio 4:1.

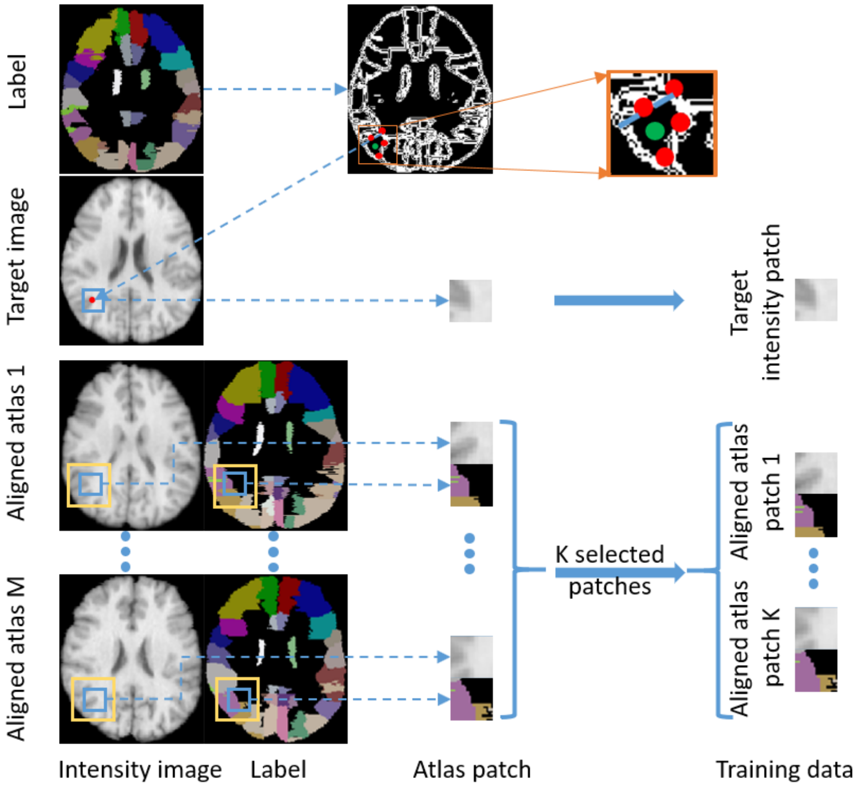

Figure 2: Flowchart illustrating *patch sampling* and *similar atlas patches selection*. (Top) We sample patches both around the boundary (e.g., red dots) and inside (e.g., green dot) the target anatomical regions of interest. (Bottom) The blue box represents a selected patch and the yellow box delineates its corresponding search neighborhood. For each target intensity patch, we identify its K most similar atlas patches. Then, each selected intensity atlas patch is coupled with its corresponding label patch to make up the training atlas data (paired with the target training patch).

3.1.2. Candidate atlas patch selection

An atlas set A contains M atlases, which is defined as $A = \{I_{A(i)}, L_{A(i)} | i = 1,2, \dots, M\}$, where $I_{A(i)}$ and $L_{A(i)}$ represent the i-th atlas intensity image and its corresponding atlas label map, respectively. For convenience, the atlas set is represented as Ω, where $\Omega = \{1,2, \dots, M\}$. A target image set B contains N samples, each defined as follows: $B_i = \{I_{B(i)}, L_{B(i)} | i = 1,2, \dots, N\}$, where $I_{B(i)}$ and $L_{B(i)}$ represent the j-th training intensity image and its corresponding label map, respectively. For each target patch $I_{B(i)}^{j}$ centered at location j, the most similar atlas intensity patches are extracted from each atlas $I_{A(i)}$ within a search neighborhood $N(j)$ based on a predefined image similarity measure. As shown in Equation 1 below, $\hat{P}$ is the collection of selected candidate atlas patches from all existing atlases. $P_{A(m)}^{n} = \{I_{A(m)}^{n}, L_{A(m)}^{n}\}$ denotes the selected label and intensity patches from atlas $m$at location n, and $I_{A(m)}^{n}$, $L_{A(m)}^{n}$ denote the intensity and label patches, respectively. $||\cdot||_2$ is the Euclidean distance.

$$\hat{P} = \{P_{A(m)}^{n}, m \in \Omega \mid \min_{n \in N(j)} ||I_{B(i)}^{j} - I_{A(m)}^{n}||_2\} \quad (1)$$

To reduce the computational time of our model, we divide our patch selection strategy into two steps. For each atlas image, we first extract their atlas patches within the first search window (with the same center location as the intensity patch and spaced out by a step size of 2 voxels). Among these patches, we find the candidate patch that has the highest similarity with the intensity patch. Then, we set up the second search window (with the same center location as the aforementioned candidate patch and spaced out by a step size of 1 voxels), and reselect the candidate patch following the same criterion, and within that new search region. Note that, to use our method on different datasets, all brain MR data are first normalized within a fixed intensity range [0, 255] using Min-Max normalization strategy before performing atlas patch selection. For example, in our validation datasets, image intensity of LONI dataset falls within a range of [0, 3000], while image intensity of SATA dataset falls within a range of [0, 4000]. We suppress the intensity value to the 85% of the max intensity value of the input image, and then normalize the image intensity value from 0 to 255. We should also note that the range [0, 255] is not very important. We have also normalized the MR data using [0, 1] and [-0.5, 0.5] intervals respectively, which did not affect the labeling performance when using a normalization interval of [0, 255]. Next, we identify the set of most similar atlas intensity patches to the target intensity patch using the Euclidean distance as follows:

$$\bar{P} = \{P_{A(m)}^{n}, m \in R, |R| = K \big| ||I_{B(i)}^{j} - I_{A(m)}^{n}||_2 \leq ||I_{B(i)}^{j} - I_{A(t)}^{n}||_2; I_{A(m)}^{n}, I_{A(t)}^{n} \in \hat{P}; t \in \Omega - R\} \quad (2)$$

By ranking all selected atlas image patches $\hat{P}$, the top K most similar patches $\bar{P}$ can be selected from the M similar patches using Equation 2.Then, the training patch $I_{B(i)}^{j}$ and its K selected atlas image patches are combined as joint input to our proposed model. R is a subset of Ω, which contains the indices of the final selected similar atlases. $|R|$ denotes the cardinal of R.

Figure 2 shows both *patch sampling* and *similar atlas patches selection* steps. In the sampling step, we extract many patches around the ROI boundary (red points) and fewer patches inside the target ROI (green point).

3.2. Multi-atlas Guided Fully Convolutional Networks (MA-FCN)

The flowchart of our proposed framework is summarized in Figure 3, which comprises three components: 1) *atlas-unique pathway*, 2) *target-patch pathway,* and 3) *atlas-aware fusion pathway*. For each candidate atlas patch, it is concatenated with the target patch to propagate independently using an *atlas-unique pathway*. On the other hand, an *atlas-aware-fusion* pathway is proposed to merge separate atlas pathways into the *target-patch* pathway. In particular, the *target-patch pathway* propagates the target patch along with the fused atlas intensity and label patches to get the final label map. Note that each training patch propagates *not only* using an independent path (*target-patch* pathway), *but also* along the *atlas-unique* pathway as it concatenates with the selected candidate atlas patch. We detail each of these three components in Sections 3.2.1, 3.2.2 and 3.2.3, respectively.

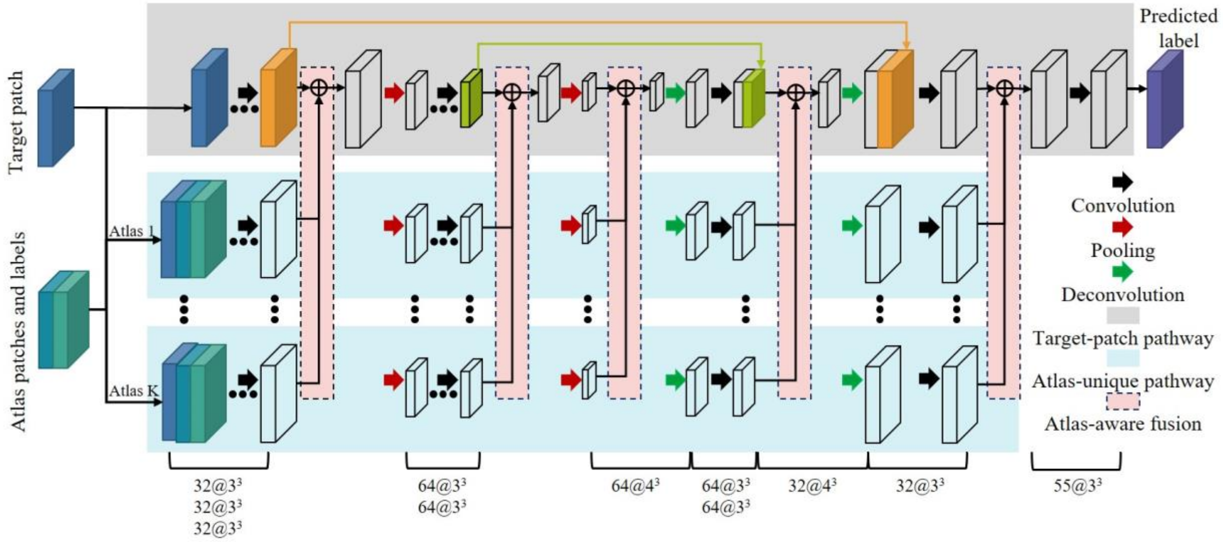

Figure3: The flowchart of the proposed Multi-Atlas Fully-Convolution Network (MA-FCN). The three pathways in MA-FCN are highlighted in gray, cyan, and pink bands. The batch normalization layer and the ReLU layer are each followed by the convolution and deconvolution layers. The symbol ⊕ denotes the concatenation of all the data together and then being convolved by a $1 \times 1 \times 1$ kernel. The parameters under the figure are the parameters of the single pathway.

3.2.1 Atlas-unique pathway

The atlas-unique pathway is designed based on the fully convolutional network (FCN), which aims to convert the atlas information (intensity and label) into comprehensive features to enhance the discrimination capacity of the model. In our previous work [54], we concatenated the atlas image and the target image together directly as input to the neural network, in order to learn the mapping from intensity image to the label map. In this method, we adopt a patch-wise ‘atlas and target’ integration strategy, where the atlas patch is treated as an enhanced feature of the target patch. However, this enhanced information might misguide the learning process since the label of the selected atlas patch might not correspond well with the true label of the target patch. To tackle this issue, instead of directly combining the atlas with the target intensity patch, we design an *atlas-unique pathway* to process each atlas patch independently.

For each atlas-unique pathway, we concatenate the target intensity patch and the atlas patch (i.e., intensity and label atlas patches) together as input to our FCN. The reason for adding atlas label patch is that the label represents strong semantic information, which can better guide the learning process. An example of the atlas-unique pathway is highlighted in cyan band in Figure 3. The structure of each atlas-unique pathway is an FCN. In the proposed model, we have several atlas-unique pathways, each processing a single atlas patch. Note that all pathways are processed independently and the weights between different pathways are not shared. The reason for designing the model in such way is that we want to build the relationship between the target patch and each atlas label patch, while taking into account the fact that different atlases have different mappings between the target patch and its label patch. In the proposed model, we order the atlas patches by the decreasing similarity, where the top atlas-unique pathway includes the most similar atlas patch, and the second pathway includes the second most similar atlas patch, etc.

3.2.2 Target-patch pathway

The target-patch pathway is used to learn the features of the target patch, as shown in the gray band in Figure 3. It is designed based on a U-Net model. We select U-Net as a basic architecture in the target-patch pathway, since U-Net architecture can combine the shadow layer feature with deep layer feature. Shadow layer features can help compensate the information loss caused by max pooling operation. Moreover, the proposed architecture will fuse the atlas feature in the latter layers, so that the U-Net structure can combine pure target information (without atlas information) into the latter layer to increase the weights of target patch features.

3.2.3. Atlas-aware fusion pathway

For each atlas, we create an atlas-unique pathway, along which the atlas patches are propagated. Hence, we create multiple independent atlas-unique pathways, each associated with a single atlas. To ultimately merge all atlas features with the target image feature, an atlas-aware fusion procedure is applied in the MA-FCN by using a convolution operation. Specifically, for all the atlas-unique pathways, the feature maps in each level are concatenated together following several convolutions. Then, a convolution layer with $1 \times 1 \times 1$ kernel is used to fuse them together, which is denoted by $\oplus$ in Figure 3. As the size of convolution kernel is one, the atlas-aware fusion is similar to a weighted sum of the learned feature maps of atlases. Unlike existing methods that define the weight based on the similarity, the weights in our framework are learned automatically by the model itself. In this paper, we use atlas-aware fusion in a hierarchical manner, instead of just using it at the very end of the model in order to make full use of the image features of the model. Specifically, we use atlas-aware fusion at each image scale (e.g., preceding each pooling layer and also following each deconvolution layer). Different image scales contain different image features. For example, in the first three layers of the model, the features contain lots of original intensity related information. But after several max pooling operations, the features may contain more advanced information such as edge.

3.2.4. Loss function

In the training stage, the output of the MA-FCN is the probability map of each class of the output patch. Suppose we have N voxels, $\hat{y}(i), i = 1,2,\dots,N$, denotes the probability of voxel i. If the class label for the corresponding golden standard is u, the loss function is defined as Equation 3:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{u=1}^{C} I(y^{(i)},u)\log(\hat{y}(i)) \tag{3}$$

Where $I(y^{(i)},u)$ means the similarity between $y^{(i)}$ and u. $I(y^{(i)},u) = \begin{cases} 0 & y^{(i)} \neq u \\ 1 & y^{(i)} = u \end{cases}$, and $y^{(i)}$ is the predicted label value. We use stochastic gradient descent with the standard back-propagation in [52] to minimize the loss function L.

4. Experiments and Results

We evaluated the proposed method on the LONI LBPA40[1] [55] dataset and SATA MICCAI 2013 challenge dataset[2] [56]. LONI dataset and SATA dataset are the two widely-used datasets

[1] http://www.loni.ucla.edu/Atlases/LPBA40

for evaluating 2D [11, 24, 57] or 3D [22, 58, 59] labeling algorithms. They contain different anatomical regions of the brain, which can provide several ways for demonstrating the validity of our proposed method**.** Both datasets include different anatomical regions of the brain. The LONI_LPBA40 dataset contains 40 T1-weighted MR brain images with 54 manually labeled ROIs, provided by the Laboratory of Neuro Imaging (LONI) from UCLA [55]. Most of the ROIs are distributed within cortical regions of the brain. Here, we used the images and their corresponding labels in our experiments. The SATA dataset is provided by MICCAI 2013 segmentation challenge workshop, in which 35 subjects (each with both intensity image and label map) are provided with 14 manually labeled ROIs. These 14 ROIs are inner regions of the brain, which cover accumbens, amygdala, caudate, hippocampus, pallidum, thalamus and putamen on both hemispheres. Both raw images and non-rigidly aligned images are provided by this dataset. Our goal in this section is to demonstrate the capability of our proposed framework in dealing with various challenges in brain image labeling.

We used CAFFE [60] framework to train our MA-FCN. The kernel weights were initialized by Xavier function, and stochastic gradient descent (SGD) was used for backpropagation. We set the start learning rate to 0.01 and used inverse learning policy, where gamma was set to 0.0001, momentum to 0.9, and the weight decay to 0.00005. These hyper parameters are chosen by trial and error, and we also use the training and validation errors to help infer the choice of hyper-parameters.

Our proposed method was implemented on GPU server (GeForce GTX TITAN X, RAM 12GB, 8 Intel(R) Core(TM) i7-6700K CPU@4.00GHz). For LONI dataset, the training batch size is 16, and for SATA dataset, the training batch size is 64.

We used Dice Similarity Coefficient (DSC) and Hausdorff Distances (HD) [61] to measures the degree of overlap between two ROIs for assessing the labeling accuracy. DSC is calculated using Equation 4, where $|\cdot|$ denotes the volume of an ROI, S_1, S_2 are two regions in the brain, and $\cap$ denotes the intersection operator. The Hausdorff Distance between sets A and B is calculated using Equation 5 and Equation 6, where $||a-b||$ is Euclidean distance.

$$DSC(S_1, S_2) = 2 \times |S_1 \cap S_2| / (|S_1| + |S_2|) \tag{4}$$

$$HD(A, B) = max(h(A, B), h(B, A) \tag{5}$$

[2] https://masi.vuse.vanderbilt.edu/workshop2013/index.php/Main_Page

$$h(A,B) = \max_{a \in A} \min_{b \in B} ||a - b|| \tag{6}$$

4.1. Evaluation on LONI LPBA40 dataset

Four-fold cross-validation is used to validate the proposed method. Specifically, in each experiment, one-fold (10 images) is randomly selected as atlases, two image folds are used for training, and the remaining fold is used for testing. The training patch size is $24 \times 24 \times 24$, and we select 8100 patches from each training image. We don't use data augmentation strategies such as flipping or rotating the cropped training patches. We increase the number of the data by densely cropping training patches from original MR image. Specifically, 150 patches are selected from each ROI, with 120 from ROI boundaries and 30 from the inside of each ROI. In the testing stage, to ensure that the testing patch can cover the entire image and have a sufficient overlap with the neighboring patches, the step size should be defined at least less than half the patch size; otherwise, there will be only one prediction for some locations. We sample the testing image with a fixed step size where patches are visited with a step size of 11 voxels. Since each voxel belongs to several overlapping patches, we use majority voting to get a final label value from all overlapping predicted label patches. For selecting candidate atlas patches, the size of the search neighborhood is set to 12 voxels, larger than the patch size in all three directions. Typically, the search region size is usually 1-2 times bigger than that of the patch size [9]. In our case, we chose the search region 1 time bigger than the patch size. For the LONI dataset, if we define the search region as 1 time bigger than the patch size, the computing time would be very high. So, we reduced the search region size. We had compared the similar patch selection result by 12 voxels larger and 24 voxels larger, and found that 87% of the selected locations remained unchanged. In the proposed architecture, the number of candidate atlas patches is set to K=3.

We compare our proposed method with U-Net (Ronneberger, Fischer et al. 2015) and FCN (Long, Shelhamer et al. 2015) architectures. The structure of the used U-Net is same as the target-patch pathway, which is shown in gray band in Figure 3. The structure of FCN is same as the atlas-unique pathway, which is shown in cyan band in Figure 3. For fair comparison, both the U-Net and FCN architectures share the same number of parameters in proposed structure. Specifically, in each layer, the number of the convolution kernels is 4 times the number of kernels in each pathway. Also, both models input 3D patches of the same size (without corresponding atlas patch compared with the input of MA-FCN). The hyper parameters such as learning rata, gamma, momentum, and the weight decay are set similarly to MA-FCN. We evaluated U-Net and FCN architectures on SATA dataset as baseline methods. Table 1 displays the mean and standard deviation of DSC for all 54 ROIs. The proposed method achieves 1.8% improvement over U-Net and 2.3% over FCN, respectively. For the HD, proposed model is smaller than both of them. Figure 4 displays the results of our method in comparison with the FCN and U-Net on all 54 ROIs. The symbol '+' indicates that MA-FCN has a statistically significant (p<0.05 by paired t-test) improvement compared with the conventional FCN method in 29 ROIs, while the symbol '*' indicates that MA-FCN has a statistically significant (p<0.05 by t-test) improvement compared with the U-Net in 28 ROIs. Figure 5 shows the visual comparison of the proposed MA-FCN with FCN and U-Net. The labeling result of the region inside the yellow box shows that, with the integration of multiple atlases, the labeling ability of our model is improved. In Figure 5 and 6, the labeling result produced by our proposed method is smoother than the ground truth. Since the ground truth is manually labeled, the discontinuity error might be occurred between adjacent slices. However, the smoother result is more biologically feasible, and our method has not reproduced this discontinuity error. Therefore, our labeling performance is not attributed by simple overfitting the data. Moreover, we also teste the trained model by using the training image, and achieve the labeling DSC of 84.3% on LONI dataset. This demonstrates that the labeling results are not overfitting the dataset.

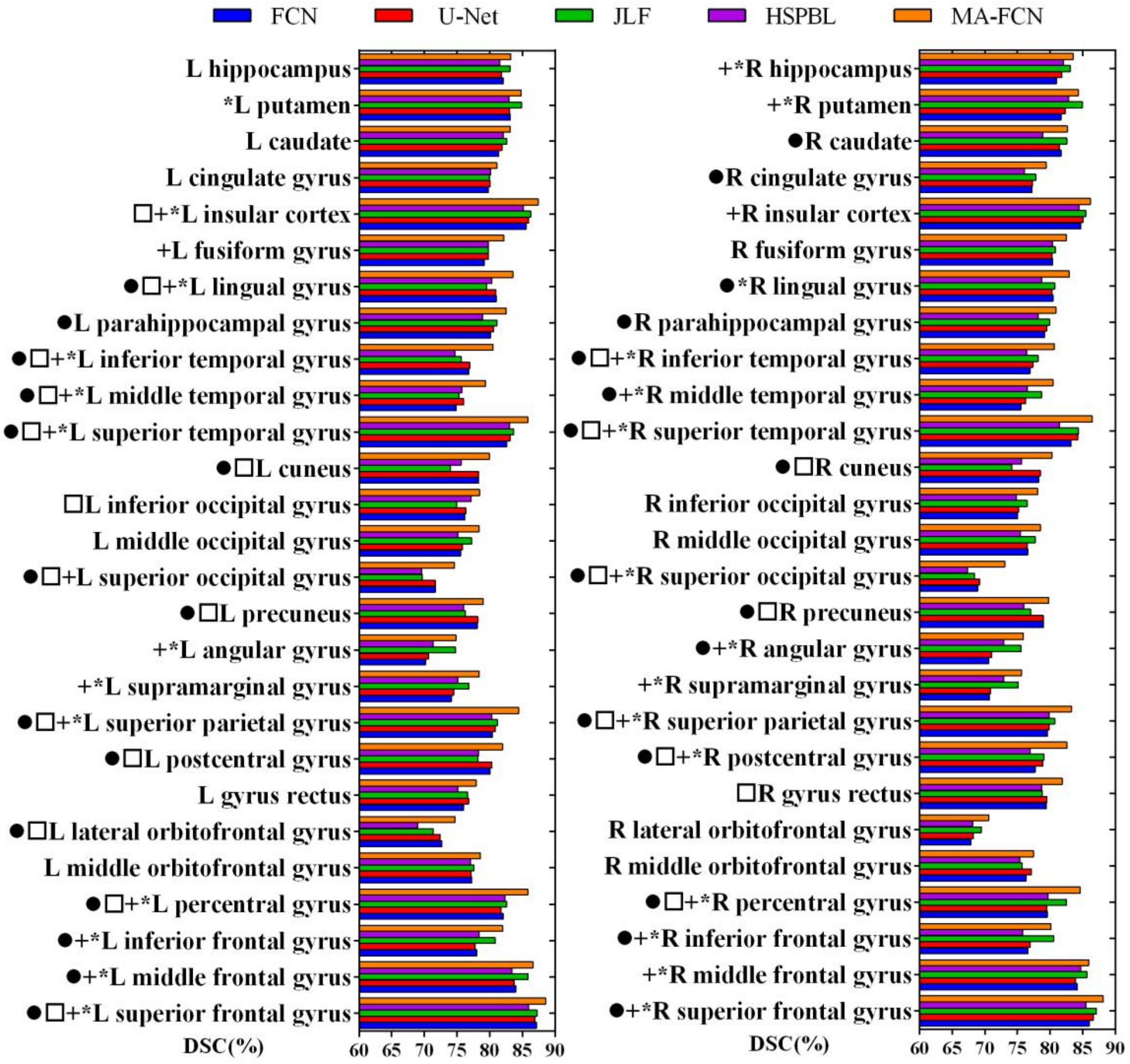

Figure 4: DSC for each ROI by FCN, U-Net, JLF, HSPBL and MA-FCN, respectively. MA-FCN outperforms both the conventional FCN and U-Net in all ROIs. The symbol '+' indicates statistically significant improvement (*p*<0.05 by paired *t*-test) *with respect to* the conventional FCN. The symbol '*' indicates statistically significant improvement (*p*<0.05 by paired *t*-test) *with respect* to U-Net. The symbol '□' indicates statistically significant improvement (*p*<0.05 by paired *t*-test) *with respect to* the JLF. The symbol '●' indicates statistically significant improvement (*p*<0.05 by paired *t*-test) *with respect to* the HSPBL.

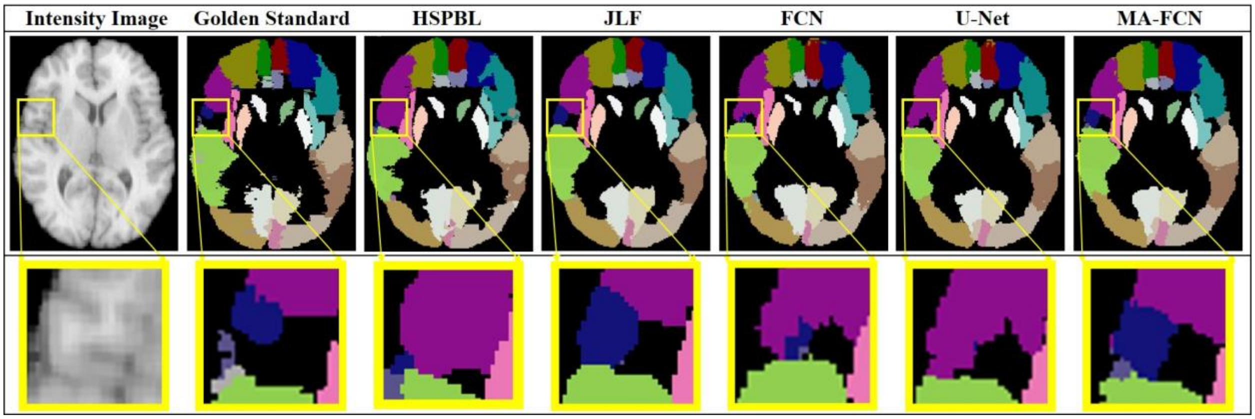

Figure 5: Visual comparison of labeling results by HSPBL, JLF, 3D patch-based FCN, U-Net, and MA-FCN for a representative subject. Our method produces more accurate labels for the regions inside the yellow box.

4.2. Evaluation on SATA MICCAI 2013 dataset

7-fold cross-validation is used in this experiment. Specifically, we divide 35 subjects into 7 groups, each group containing 5 subjects. Next, we randomly select 2 folds as atlas images, 4 folds as our training set, and the remaining fold as our test set. Since the number of ROIs to label is smaller than that in LONI dataset, we set the training patch size to $12 \times 12 \times 12$, and select 4200 patches from each training image. Note that 300 patches are selected from each ROI, including 240 around the boundary and 60 inside the ROI. We evenly visit patches with a step size of 5 voxels. For selecting the candidate atlas patches, the size of the search neighborhood is set to 12 voxels larger than the patch size in all three directions. The number of candidate atlas patches is set to K=3.

The mean and standard deviation of DSC for all comparison methods are listed in Table 1. In terms of DSC, our proposed method has a 0.8% improvement compared with U-Net and 1.2% improvement compared with FCN. The HD of the proposed model is smaller than both comparison models. Figure 6 gives visual comparison of our labeling results with the golden standard. The labeling result of the region inside the yellow box shows that, with the integration of multiple atlases, the labeling ability of our model is improved.

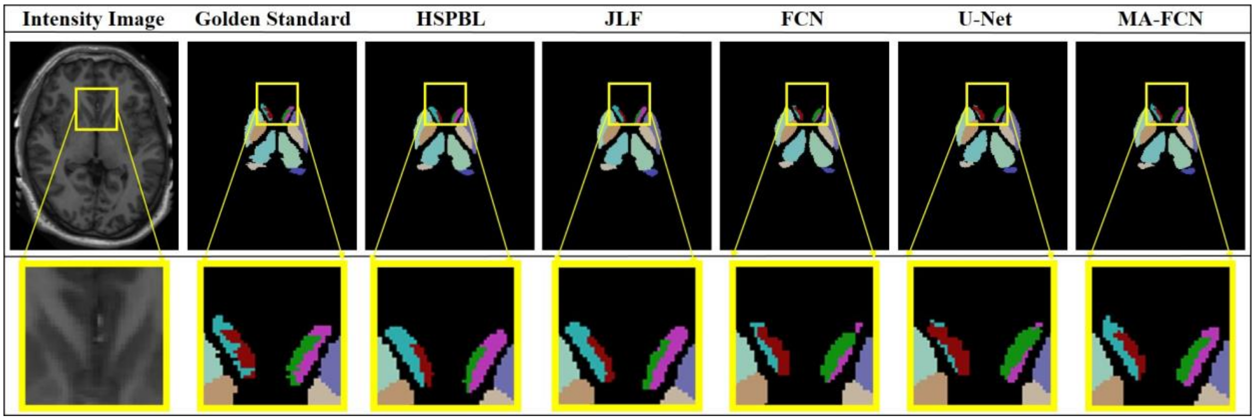

Figure 6: Visual comparison of labeling results by HSPBL, JLF, 3D patch-based FCN, U-Net, and MA-FCN for a representative subject from SATA dataset. Our method produces more accurate labels for the regions inside the yellow box.

4.3. Parameter tuning

4.3.1 Patch size

In order to evaluate the influence of the patch size on labeling ROIs with different sizes, we selected 12 representative ROIs with different volume sizes from the LONI_LPBA40 dataset and 6 representative ROIs with different volume sizes from SATA MICCAI 2013 dataset. Specifically, for LONI dataset, these ROIs include the right/left inferior frontal gyrus (IFG), right/left precentral gyrus (PG), right/left precuneus (PC), right/left para hippocampus gyrus (PHG), right/left caudate (CD) and right/left hippocampus (HC). The volumes of right/left IFG and left/right PG contain about 25,000 voxels, the volumes of right/left PC and PHG contain about 10,000 voxels, and the volumes of right/left CD and HC contain about 5,000 voxels. For SATA dataset, these ROIs include the right/left accumbens (AC), right/left caudate (CA) and right/left putamen (PU). The right/left AC contains about 500 voxels, the right/left CA contains about 3000 voxels, and the right/left PU contains about 8000 voxels.

We varied the patch size between $8 \times 8 \times 8$ and $28 \times 28 \times 28$ for the LONI dataset by 4-fold cross-validation. Figure 7 shows the labeling performance using different patch sizes. We note that the performance has been improved when increasing the patch size from 8 to 12 and then remains stable when the patch size falls between 12 and 24. However, when the patch size exceeds 24, the labeling accuracy starts to decrease. This is mainly because a small patch contains less structural information while two patches from different locations may look similar. This may cause the model to fail in distinguishing between them. Conversely, using larger patches would decrease similarity with the selected atlas patches. The larger the patch size, the more structure is included in the patch, so the dissimilarity between target patch and selected atlas patches is increased. For the target patch, the number of the wrong label will increase (if the atlas label is directly used as target patch label), thereby causing a drop in the labeling accuracy.

We also varied the patch size between 8 × 8 × 8 and 24 × 24 × 24 for the SATA dataset by 7-fold cross-validation. Figure 8 shows the labeling performance using different patch sizes. The performance increases from patch size 8 to 12 for all ROIs and keeps stable from 12 to 20 on large and mediate ROIs, but decreases in small ROIs. When the patch size keeps increasing, the labeling accuracy decreases in all ROIs. The reason that the labeling accuracy of small ROI keeps decreasing from patch size 12 is because of small size of those ROIs. If the patch size is large, those small ROIs only account for a small portion of the patch, thus causing the poor learning in these ROIs.

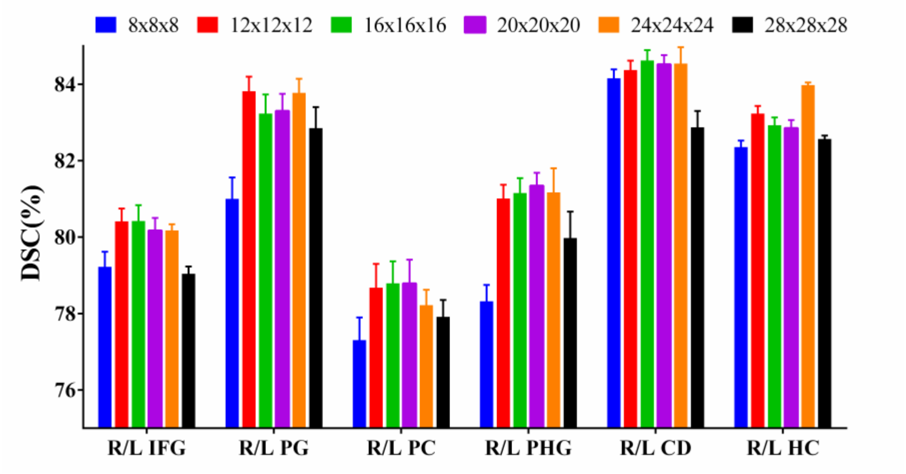

Figure 7: The influence of using different label patch sizes on labeling 12 representative ROIs on the LONI_LPBA40 dataset. By enlarging the patch size between8 × 8 × 8 and 12 × 12 × 12, the performance largely increases, and then remains stable between patch sizes of 12 × 12 × 12 and 24 × 24 × 24. As the patch size continues to increase, the performance decreases. Note that the DSC is the average value across all four-fold cross-validation.

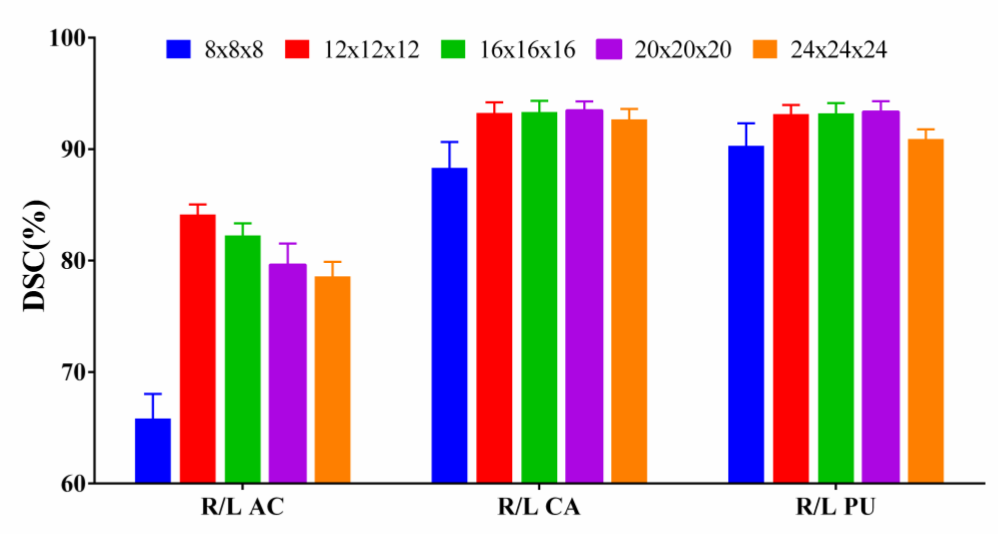

Figure 8: The influence of using different label patch sizes on labeling 6 representative ROIs on the SATA MICCAI 2013 dataset. By enlarging the patch size between $8 \times 8 \times 8$ and $12 \times 12 \times 12$, the performance largely increases on all ROIs, while remaining stable between patch sizes of $12 \times 12 \times 12$ and $20 \times 20 \times 20$on mediate and large ROIs but beginning decreasing for small ROIs. As the patch size continues to increase, the performance decreases. The DSC is the average of all the 35 testing data by seven-fold cross-validation.

4.3.2 The number of atlas-unique pathways

In the proposed method, the top K similar candidate atlas patches are selected from affine-aligned atlases as input to the atlas-unique pathways for helping improve the labeling performance. We evaluated the performance by tuning the parameter K on both LONI and SATA datasets. The value of K ranges from 0 to 4. Figure 9 shows the evaluation result *with respect to* the number of the atlas-unique pathways. We can clearly see that the performance of our model increases significantly from 0 atlas-unique pathways to 1 atlas-unique pathway, indicating that the atlas and label information did aid in boosting the labeling quality. As the number of patches increases, the labeling quality is refined, but the memory and processing time cost also increase. To balance the performance and the memory cost (and also processing time), we use 3 atlas-unique pathways in our model.

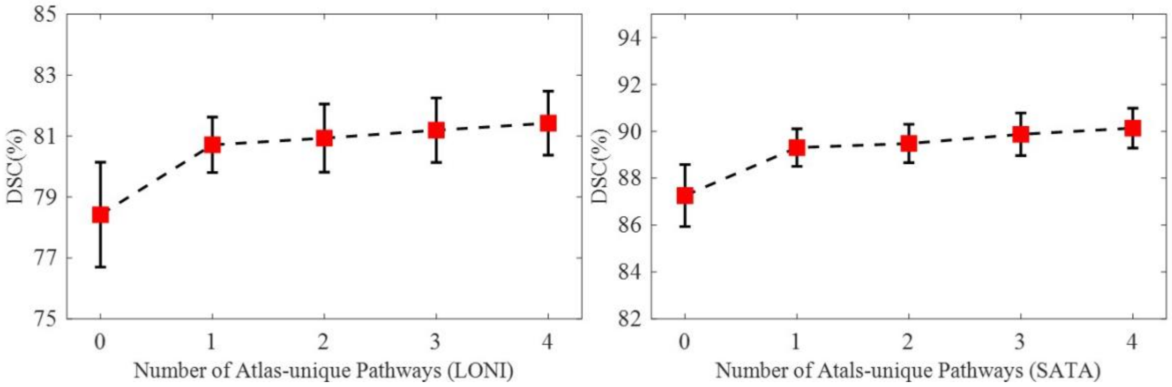

Figure 9: Evaluation on the number of atlas-unique pathways using both LONI and SATA dataset, in terms of DSC (%). The performance increases with the increase of the number of candidate atlas patches.

4.4. Comparison with state-of-the-art methods

To evaluate the labeling performance, we compare our proposed method with two state-of-the-art methods on both LONI and SATA datasets. The comparison methods include 1) HSPBL [11] and JLF [20] (antsJointFusion command in ANTs toolbox). JLF is a registration-based labeling method, and HSPBL is a patch-based labeling method. The detailed comparisons are listed in Table 1. We reproduced all results shown in Table 1. Both methods use leave-one-out strategy to evaluate all the test data and the configure parameters are same as the original papers.

For LONI dataset, our proposed MA-FCN improved the labeling accuracy by 2% in comparison with JLF. Compared with the HSPBL method, our proposed method achieves 2.72% improvement. Figure 4 displays the results of our method in comparison with the HSPBL and JLF on all 54 ROIs. The symbol '●' indicates that MA-FCN has a statistically significant (p<0.05 by paired t-test) improvement compared with the HSPBL method in 31 ROIs, while the symbol '□' indicates that MA-FCN has a statistically significant (p<0.05 by t-test) improvement compared with the JLF in 23 ROIs. Figure 5 shows the visual comparison of the proposed MA-FCN with HSPBL and JLF on LONI dataset. For SATA dataset, our proposed MA-FCN improved the labeling accuracy by 1.81% in comparison with JLF and 2.91% more than the HSPBL method. For the Hausdorff distance, our method has the smallest value for both datasets. Figure 6 gives visual comparison of our labeling results with the HSPBL and JLF on SATA dataset.

Table 1. Comparison with state-of-the-art methods on two datasets.

LONI LPBA40					
Method	HSPBL	JLF	FCN	U-Net	MA-FCN
HD(voxel)	22.95±4.81	17.59±**3.14**	21.50±4.69	16.25±4.00	**14.11**±3.22
DSC(%)	78.47±2.33	79.19±**0.98**	78.88±1.07	79.42±1.12	**81.19**±1.06

SATA					
Method	HSPBL	JLF	FCN	U-Net	MA-FCN
HD(voxel)	4.18±1.73	3.84±1.30	3.34±0.92	2.76±0.81	**2.38±0.71**
DSC(%)	86.13±2.75	87.23±1.91	87.82±1.37	88.25±1.42	**89.04±1.30**

The average testing time is 7 minutes for each subject. In particular, 5 minutes are used for preparing the test patches on CPU and about 2 minutes used for inferencing the test patches by the trained model on the GPU platform. For the registration-based method [20], the average labeling time for one subject is 120 minutes on CPU. Our proposed method is much faster than registration-based method. For the patch-based method [11], the labeling time is 40 minutes. Notably, our method is faster. For example, for ConvNet-based methods, the average labeling time is 2 minutes. On the other hand, although ConvNet-based methods are faster than MA-FCN, MA-FCN can achieve higher labeling accuracy, as indicated in Section 4.1. The specific time usage and memory cost is listed in Table 2. The sign "-" means no this step in the method.

Table 2. The comparison of time usage and memory cost for different methods

	Affine reg.	Deform reg.	Patch selection	Label fusion	Inference	Training
Memory	<1G	<1G	<1G	3G	1G	12G
	CPU	CPU	CPU	CPU	GPU	GPU
HSPBL	8 min (4 threads)	240 min (4 threads)	-	40 min	-	-
JLF	8 min (4 threads)	240 min (4 threads)	-	120 min	-	-
FCN	-	-	-	-	90 s	12 h
U-Net	-	-	-	-	90 s	14 h
MA-FCN	8 min (4 threads)	-	5 min (2 threads)	-	140 s	20 h

5. Discussion

In this paper, we proposed an automated labeling framework of brain images, by integrating multiple-atlas based labeling approaches into an FCN architecture. Previously, several neural network-based methods aimed to integrate data from multiple sources or different modalities by concatenating them together for network training [54, 62-64]. Our proposed MA-FCN falls into the same category, but it has more appealing aspects. For instance, Fang *et al.* [54] simply concatenate the training patch, atlas intensity patches, and label maps together as inputs to the U-Net, whereas the atlas information is propagated independently and fused together in our MA-FCN architecture.

The proposed MA-FCN outperformed U-Net [54] as it increased the labeling accuracy by 0.8%. We note that atlas label patches are selected from the atlas, not from the target image, hence the label values might not perfectly match with the ground-truth label of the target patch. To address this issue, we defined the *atlas-unique pathway* in our FCN, where label information can be propagated independently. Guided by the ground truth, the label can be refined by the convolution operation. Then, the refined label maps are fused into target patch to get the final label maps.

The label map is a strong semantic information that is leveraged and integrated into our proposed deep learning architecture. Both the feature information from the *target-patch pathway* and the *atlas-unique pathway* make contributions to the labeling works in the MA-FCN. Here, we further validate their importance in the framework, by conducting a labeling experiment using our proposed method without the *target-patch pathway*, and leaving only the *atlas-aware fusion* and the *atlas-unique pathways*. The labeling performance for the LONI-LBPA 40 is reduced to 76.91 ± 1.21%, compared with the MA-FCN method with all three components included (81.19±1.06%) as shown in Table 1. Meanwhile, the labeling performance for U-Net FCN is 79.42±1.12%, which can also be considered as the MA-FCN method using only the component of *target-patch pathway*. Therefore, this experiment validates that all three components help improve the labeling performance for the MA-FCN method.

In Rousseau et al. [28], they found that accurate correspondences derived from non-rigid registration could improve the labeling performance. Here, we evaluate the performance of our proposed architecture by replacing the affine registration with non-rigid registration. For the SATA dataset, the organizer had already provided non-rigid registration results. For the LONI dataset, we use SyN registration method integrated in ANTs software to non-rigidly register atlases to the target image. The DSC on SATA dataset is 89.27±1.07%, and the performance on LONI dataset is 81.81%. These results show that non-rigid registration can slightly improve the label performance of our proposed architecture than affine registration.

Despite its appealing aspects, our MA-FCN method is limited by a large memory cost when compared with the conventional FCN and U-Net architectures. Although the added similar atlas patches improve the labeling performance, the memory cost increases largely. For example, the memory cost is almost two times the ordinary FCN for a MA-FCN with three pathways. Moreover, even though our MA-FCN method needs fewer iterations to converge, the training time for each iteration increase as the complexity of network architecture increases, which leads to a longer training time. Future work will focus on how to reduce the parameters of the network. Alternatively, we will consider using ResNet [65, 66] structure as a backbone structure in our MA-FCN method. ResNet structure is proved to be more efficient and uses less memory than the general convolutional network.

6. Conclusion

In this work, we have proposed a novel multi-atlas guided fully convolutional networks (MA-FCN) for brain labeling. Different from conventional ConvNet methods, we integrated atlas intensity and label information through new pathways embedded in the proposed FCN architecture. The MA-FCN contains three propagation pathways: *atlas-unique pathway, atlas-aware fusion pathway*, and *target-patch pathway*. The *atlas-uniquepathway* can amend the wrong labels in the atlas by using the convolution operation. The *atlas-aware fusion pathway* gives each voxel in the candidate atlas patch a weight and fuses them together at the voxel level. Last, the *target-patch pathway* propagates the target patch and the fused information. In this way, MA-FCN combines the advantages of both multi-atlas-based and ConvNet labeling methods. Our method does not require non-rigid registration, but it can still achieve better or comparable results with the state-of-the-art multi-atlas-based methods on LONI dataset and much better performance on SATA dataset. Moreover, the idea of our proposed architecture can also be easily applied to other ConvNet methods such as RNN [67] or LSTM [68].

Acknowledgement

This work was supported in part by The National Key Research and Development Program of China (2017YFB130202) and National Natural Science Foundation of China (91520202, 81701785), Youth Innovation Promotion Association CAS (2012124), the CAS Scientific Research Equipment Development Project (YJKYYQ20170050) and the Beijing Municipal Science & Technology Commission (Z181100008918010). This work was also supported by

NIH grants (EB006733, EB008374, MH100217, MH108914, AG041721, AG049371, AG042599, AG053867, EB022880, MH110274). Dr. S.-W. Lee was partially supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451).

References

1. Chen, X., et al., *Extraction of dynamic functional connectivity from brain grey matter and white matter for MCI classification.* Human brain mapping, 2017. **38**(10): p. 5019-5034.
2. Zhou, J., et al., *Predicting regional neurodegeneration from the healthy brain functional connectome.* Neuron, 2012. **73**(6): p. 1216-1227.
3. Holland, D., et al., *Structural growth trajectories and rates of change in the first 3 months of infant brain development.* JAMA neurology, 2014. **71**(10): p. 1266-1274.
4. Bullmore, E.T. and D.S. Bassett, *Brain graphs: graphical models of the human brain connectome.* Annu Rev Clin Psychol, 2011. **7**: p. 113-40.
5. Liu, L., et al., *Altered cerebellar functional connectivity with intrinsic connectivity networks in adults with major depressive disorder.* PLoS One, 2012. **7**(6): p. e39516.
6. Ingalhalikar, M., et al., *Sex differences in the structural connectome of the human brain.* Proceedings of the National Academy of Sciences, 2014. **111**(2): p. 823-828.
7. Zhang, L., et al., *Learning-based structurally-guided construction of resting-state functional correlation tensors.* Magnetic resonance imaging, 2017. **43**: p. 110-121.
8. Langerak, T.R., et al., *Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE).* IEEE transactions on medical imaging, 2010. **29**(12): p. 2000-2008.
9. Coupé, P., et al., *Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation.* NeuroImage, 2011. **54**(2): p. 940-954.
10. Ma, G., et al., *Nonlocal atlas‐guided multi‐channel forest learning for human brain labeling.* Medical physics, 2016. **43**(2): p. 1003-1019.
11. Wu, G., et al., *Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition.* NeuroImage, 2015. **106**: p. 34-46.
12. Tong, T., et al., *Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling.* NeuroImage, 2013. **76**: p. 11-23.
13. Sanroma, G., et al., *A transversal approach for patch-based label fusion via matrix completion.* Medical image analysis, 2015. **24**(1): p. 135-148.
14. Zhang, J., et al., *Brain atlas fusion from high-thickness diagnostic magnetic resonance images by learning-based super-resolution.* Pattern recognition, 2017. **63**: p. 531-541.
15. Wu, G., et al., *A generative probability model of joint label fusion for multi-atlas based brain segmentation.* Medical image analysis, 2014. **18**(6): p. 881-890.
16. Shen, D. and C. Davatzikos, *HAMMER: hierarchical attribute matching mechanism for elastic registration.* IEEE transactions on medical imaging, 2002. **21**(11): p. 1421-1439.
17. Klein, A., et al., *Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration.* Neuroimage, 2009. **46**(3): p. 786-802.
18. Kim, M., et al., *Automatic hippocampus segmentation of 7.0 Tesla MR images by combining multiple atlases and auto-context models.* NeuroImage, 2013. **83**: p. 335-345.
19. Giraud, R., et al., *An optimized patchmatch for multi-scale and multi-feature label fusion.* NeuroImage, 2016. **124**: p. 770-782.
20. Wang, H., et al., *Multi-atlas segmentation with joint label fusion.* IEEE transactions on pattern analysis and machine intelligence, 2013. **35**(3): p. 611-623.
21. Iglesias, J.E. and M.R. Sabuncu, *Multi-atlas segmentation of biomedical images: a survey.* Medical image analysis, 2015. **24**(1): p. 205-219.
22. Tu, Z. and X. Bai, *Auto-context and its application to high-level vision tasks and 3d brain image segmentation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010. **32**(10): p. 1744-1757.
23. Hao, Y., et al., *Local label learning (LLL) for subcortical structure segmentation: application to hippocampus segmentation.* Human brain mapping, 2014. **35**(6): p. 2674-2697.
24. Zikic, D., B. Glocker, and A. Criminisi, *Encoding atlases by randomized classification forests for efficient multi-atlas label propagation.* Medical image analysis, 2014. **18**(8): p. 1262-1273.

25. Pereira, S., et al., *Automatic brain tissue segmentation in MR images using random forests and conditional random fields*. Journal of neuroscience methods, 2016. **270**: p. 111-123.

26. Khalifa, F., et al. *A random forest-based framework for 3D kidney segmentation from dynamic contrast-enhanced CT images*. in *Image Processing (ICIP), 2016 IEEE International Conference on*. 2016. IEEE.

27. Zhang, L., et al., *Concatenated spatially-localized random forests for hippocampus labeling in adult and infant MR brain images*. Neurocomputing, 2017. **229**: p. 3-12.

28. Rousseau, F., P.A. Habas, and C. Studholme, *A supervised patch-based approach for human brain labeling*. IEEE transactions on medical imaging, 2011. **30**(10): p. 1852-1862.

29. Zhang, L., et al., *Automatic labeling of MR brain images by hierarchical learning of atlas forests*. Medical physics, 2016. **43**(3): p. 1175-1186.

30. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015. Springer.

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

32. Badrinarayanan, V., A. Kendall, and R. Cipolla, *Segnet: A deep convolutional encoder-decoder architecture for scene segmentation*. IEEE transactions on pattern analysis and machine intelligence, 2017.

33. Chen, L.-C., et al., *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*. arXiv preprint arXiv:1606.00915, 2016.

34. Li, C. and M. Wand. *Combining markov random fields and convolutional neural networks for image synthesis*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

35. Nie, D., et al. *Medical image synthesis with context-aware generative adversarial networks*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2017. Springer.

36. Van Nguyen, H., K. Zhou, and R. Vemulapalli. *Cross-domain synthesis of medical images using efficient location-sensitive deep network*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015. Springer.

37. Long, J., E. Shelhamer, and T. Darrell. *Fully convolutional networks for semantic segmentation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

38. Nie, D., et al. *Fully convolutional networks for multi-modality isointense infant brain image segmentation*. in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. 2016. IEEE.

39. Wolz, R., et al., *LEAP: learning embeddings for atlas propagation*. NeuroImage, 2010. **49**(2): p. 1316-1325.

40. Jia, H., P.-T. Yap, and D. Shen, *Iterative multi-atlas-based multi-image segmentation with tree-based registration*. NeuroImage, 2012. **59**(1): p. 422-430.

41. Rohlfing, T., et al., *Quo vadis, atlas-based segmentation?*, in *Handbook of biomedical image analysis*. 2005, Springer. p. 435-486.

42. Artaechevarria, X., A. Munoz-Barrutia, and C. Ortiz-de-Solórzano, *Combination strategies in multi-atlas image segmentation: Application to brain MR data*. IEEE transactions on medical imaging, 2009. **28**(8): p. 1266-1277.

43. Isgum, I., et al., *Multi-atlas-based segmentation with local decision fusion—Application to cardiac and aortic segmentation in CT scans*. IEEE transactions on medical imaging, 2009. **28**(7): p. 1000-1010.

44. Rohlfing, T., et al., *Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains*. NeuroImage, 2004. **21**(4): p. 1428-1442.

45. Sabuncu, M.R., et al., *A generative model for image segmentation based on label fusion*. IEEE transactions on medical imaging, 2010. **29**(10): p. 1714-1729.

46. Warfield, S.K., K.H. Zou, and W.M. Wells, *Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation*. IEEE transactions on medical imaging, 2004. **23**(7): p. 903-921.

47. Zhan, Y. and D. Shen. *Automated segmentation of 3D US prostate images using statistical texture-based matching method*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2003. Springer.

48. Sanroma, G., et al., *Learning to rank atlases for multiple-atlas segmentation*. IEEE transactions on medical imaging, 2014. **33**(10): p. 1939-1953.

49. Zhang, D., et al., *Sparse patch-based label fusion for multi-atlas segmentation*. Multimodal brain image analysis, 2012: p. 94-102.

50. Zhang, W., et al., *Deep convolutional neural networks for multi-modality isointense infant brain image segmentation.* NeuroImage, 2015. **108**: p. 214-224.

51. Havaei, M., et al., *Brain tumor segmentation with deep neural networks.* Medical image analysis, 2017. **35**: p. 18-31.

52. LeCun, Y., et al., *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 1998. **86**(11): p. 2278-2324.

53. Smith, S.M., et al., *Advances in functional and structural MR image analysis and implementation as FSL.* Neuroimage, 2004. **23**: p. S208-S219.

54. Fang, L., et al. *Brain Image Labeling Using Multi-atlas Guided 3D Fully Convolutional Networks*. in *International Workshop on Patch-based Techniques in Medical Imaging*. 2017. Springer.

55. Shattuck, D.W., et al., *Construction of a 3D probabilistic atlas of human cortical structures.* Neuroimage, 2008. **39**(3): p. 1064-1080.

56. Bennett Landman, S.W., *2013 Diencephalon Free Challenge*. 2013.

57. Bao, S. and A.C. Chung, *Multi-scale structured CNN with label consistency for brain MR image segmentation.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2018. **6**(1): p. 113-117.

58. Bao, S., et al., *3D Randomized Connection Network With Graph-Based Label Inference.* IEEE Transactions on Image Processing, 2018. **27**(8): p. 3883-3892.

59. Wu, Z., et al., *Robust brain ROI segmentation by deformation regression and deformable shape model.* Medical image analysis, 2018. **43**: p. 198-213.

60. Jia, Y., et al. *Caffe: Convolutional architecture for fast feature embedding*. in *Proceedings of the 22nd ACM international conference on Multimedia*. 2014. ACM.

61. Taha, A.A. and A. Hanbury, *Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool.* BMC medical imaging, 2015. **15**(1): p. 29.

62. Xiang, L., et al., *Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI.* Neurocomputing, 2017. **267**: p. 406-416.

63. Rohé, M.-M., et al. *SVF-Net: Learning Deformable Image Registration Using Shape Matching*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2017. Springer.

64. Yang, X., R. Kwitt, and M. Niethammer, *Quicksilver: Fast Predictive Image Registration-a Deep Learning Approach.* arXiv preprint arXiv:1703.10908, 2017.

65. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

66. Szegedy, C., et al. *Inception-v4, inception-resnet and the impact of residual connections on learning*. in *AAAI*. 2017.

67. Graves, A., et al. *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*. in *Proceedings of the 23rd international conference on Machine learning*. 2006. ACM.

68. Stollenga, M.F., et al. *Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation*. in *Advances in neural information processing systems*. 2015.