Neural Networks 116 (2019) 25–34

Contents lists available at ScienceDirect

Neural Networks 小顶刊

journal homepage: www.elsevier.com/locate/neunet

Semi-supervised deep learning of brain tissue segmentation

Ryo Ito [a], Ken Nakae [a], Junichi Hata [b,c], Hideyuki Okano [b,c], Shin Ishii [a,*]

[a] Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Kyoto, 606-8501, Japan
[b] Department of Physiology, Keio University School of Medicine, Shinjuku-ku, Tokyo, 160-8582, Japan
[c] Laboratory for Marmoset Neural Architecture, RIKEN Center for Brain Science, Wako-shi, Saitama, 351-0198, Japan

分割脑区
半监督用的标签少，标签全用效果更好

ARTICLE INFO

Article history:
Received 3 April 2018
Received in revised form 28 September 2018
Accepted 22 March 2019
Available online 1 April 2019

Keywords:
Brain tissue segmentation
Semi-supervised learning
Image registration
Deep neural network

ABSTRACT

Brain image segmentation is of great importance not only for clinical use but also for neuroscience research. Recent developments in deep neural networks (DNNs) have led to the application of DNNs to brain image segmentation, which required extensive human annotations of whole brain images. Annotating three-dimensional brain images requires laborious efforts by expert anatomists because of the differences among images in terms of their dimensionality, noise, contrast, or ambiguous boundaries that even prevent these experts from necessarily attaining consistency. This paper proposes a semi-supervised learning framework to train a DNN based on a relatively small number of annotated (labeled) images, named atlases, but also a relatively large number of unlabeled images by leveraging image registration to attach pseudo-labels to images that were originally unlabeled. We applied our proposed method to two different datasets: open human brain images and our original marmoset brain images. When provided with the same number of atlases for training, we found our method achieved superior and more stable segmentation results than those by existing registration-based and DNN-based methods.

对三维大脑图像进行注释需要专业解剖学家的艰苦努力，因为图像在维度、噪声、对比度或模糊边界方面存在差异，甚至妨碍了这些专家获得一致性

1. Introduction

Segmentation of brain images plays an important role not only in clinical diagnosis to help assess neurological diseases but also in basic neuroscience research. In brain image segmentation, given a brain image typically acquired by magnetic resonance imaging (MRI), we estimate an annotated (labeled) image, which is categorized into several anatomical/structural regions of which the set has been prepared a priori for every voxel. The segmentation provides a quantitative evaluation of brain tissue volumetry, and it enables objective diagnosis and research rather than visual inspection by experts. The volume change in some brain regions can be used as an important biomarker; for example, neurodegenerative diseases such as Alzheimer's disease are known to be associated with shrinkage of some brain regions (Giorgio & De Stefano, 2013). Basic neuroscience research such as connectomics also demands segmentation for its pre- or post-processing (Smith, Tournier, Calamante, & Connelly, 2012).

The most straightforward way to segment brain images is to manually annotate every voxel of a brain image. In reality, however, this approach is difficult because of the presence of noise and/or differences in the contrast between brain images. Apart from this, the appearance of some boundaries between different brain regions varies, thereby complicating the annotation even by expert anatomists due to the lack of consistent criteria. Moreover, because brain images are three-dimensional (3D), it is laborious to attach labels in a voxel-wise manner (Hanbury, 2008). In contrast, simply acquiring brain images is relatively easy; there are thousands or tens of thousands of brain images available, for example, from Human Connectome Project.[1] In view of the above, automatic brain segmentation techniques have started attracting much attention as neuroscience enters the era of big data.

3D分割脑组织困难

Image registration has been used to automatically segment brain images (Cabezas, Oliver, Lladó, Freixenet, & Cuadra, 2011). This technique estimates the spatial correspondence between a manually labeled 3D brain image, known as an atlas, and another 3D brain image (target brain) that needs to be segmented. After registration, the label information of the atlas is transferred to the target brain according to the estimated correspondence. This method is advantageous because the registration process often effectively preserves the local continuity of the two brains, hence preserving the topological structure of the segmented regions in the target brain. On the other hand, the method may generate some segmentation errors especially near the boundaries between regions, because it only relies on the structural similarity between the two brains. In addition, image registration

* Corresponding author.
E-mail address: ishii@i.kyoto-u.ac.jp (S. Ishii).

[1] Human Connectome Project: http://www.humanconnectomeproject.org/.

https://doi.org/10.1016/j.neunet.2019.03.014
0893-6080/© 2019 Elsevier Ltd. All rights reserved.

usually estimates the correspondence by using an iterative approach to achieve 3D deformation, which is computationally very expensive.

深度学习在图像处理中的优势

In recent years, a number of segmentation methods incorporating machine-learning-based image processing (Ashburner & Friston, 2005; Schnell, Saur, Kreher, Hennig, Burkhardt, & Kiselev, 2009; Zikic, Glocker, & Criminisi, 2013), especially deep neural networks (DNNs), has been reported (de Brébisson & Montana, 2015; Moeskops, Viergever, Mendrik, de Vries, Benders, & Išgum, 2016). The approach involves training a DNN based on a large set of human annotated atlases, after which the DNN is applied to the brain images for segmentation. Although training the DNN is time consuming, segmentation based on the trained DNN is computationally efficient, because of its feedforward architecture. Another advantage of DNNs is their generalization ability; owing to their architecture incorporating hierarchically arranged convolution and pooling layers, they are robust against shift/rotation and blurring in the given image. Owing to these effective characteristics, segmentation based on DNNs has proven to be superior to other methods based on conventional machine learning (Chen, Dou, Yu, Qin, & Heng, 2017; Litjens, Kooi, Bejnordi, Setio, Ciompi, Ghafoorian, van der Laak, van Ginneken, & Sánchez, 2017; Zhang, Li, Deng, Wang, Lin, Ji, & Shen, 2015).

提出了一种半监督学习方法，它试图训练一个DNN基于一个相对较小的注释（标记）图谱集和一个相对较大的未标记脑图像集

In this study, we propose a semi-supervised learning approach, which attempts to train a DNN based on a relatively small set of annotated (labeled) atlases and a relatively large set of unlabeled brain images. Using image registration between the atlases and other unlabeled images, we attach a pseudo-label to every voxel in the unlabeled images. After this step, we train a DNN based on a combined dataset of the atlases and pseudo-annotated brain images. However, this naïve idea is not necessarily effective, because the data-augmented dataset for training a DNN should include label errors stemming from unsatisfactory image registration. It should be noted that if image registration was perfect, there would be no need to train a DNN for segmentation, because registration-based segmentation, which we call label propagation (LP), would also be expected to function in a perfect manner. We overcome the problem associated with erroneous labeling by employing a probabilistic model in which the true label of an originally unlabeled image is assumed to be an unobservable (hence a hidden) variable, and the pseudo-label attached by image registration is probabilistically observed by adding spatial noise to the true label. We train this probabilistic model by incorporating a DNN structural model using an expectation–maximization (EM) algorithm and estimate the true label image and the parameters of the DNN simultaneously. The new method is designed to recover the incorrect labels attached to the originally unlabeled images within the E-step of the EM algorithm.

Our new method is evaluated by applying it to open benchmark human images registered in the Internet Brain Segmentation Repository (IBSR) and our original marmoset brain image dataset acquired for the Brain/MINDS project (Okano, Miyawaki, & Kasai, 2015; Okano, Sasaki, Yamamori, Iriki, Shimogori, Yamaguchi, Kasai, & Miyawaki, 2016). In comparison with existing registration-based and DNN-based methods, the proposed method showed higher and more stable segmentation accuracies than existing methods, when using the same number of labeled images.

2. Related work

2.1. Image registration

Given a set of a reference image I (here, labeled) and a source image J (here, unlabeled), image registration estimates spatial transformation $\varphi : \Omega \to \Omega$ that satisfies

$$I(\varphi^{-1}(\mathbf{x})) \approx J(\mathbf{x}) \quad (\forall \mathbf{x} \in \Omega), \tag{1}$$

where each image is a function from each voxel to its set of intensity $I : \Omega \to \mathcal{R}^K$, where K is the number of the images using our analysis, and $\Omega \subset \mathcal{R}^3$ is the 3D region on which the image voxels are defined. Let the reference image I be associated with a labeled image $L_I : \Omega \to \{0, 1\}^C$, annotated, i.e., labeled, by an expert, where C is the number of labels. According to the label propagation, we estimate the labeled image L_J of the unlabeled source image J:

$$L_J(\mathbf{x}) \leftarrow L_I(\varphi^{-1}(\mathbf{x})). \tag{2}$$

Among many existing image registration methods, large deformation models such as a symmetric image normalization method (SyN) (Avants, Epstein, Grossman, & Gee, 2008) achieved one of the highest performances on human brain datasets (Klein, Andersson, Ardekani, Ashburner, Avants, Chiang, Christensen, Collins, Gee, Hellier, et al., 2009). They provide us with diffeomorphic transformation which transforms a reference image I to a source image J. Diffeomorphic mapping has several advantages; it involves an inversion of the transformation, which can be used to transform the source image to the reference image in an inverse manner. In addition, owing to the continuous application of deformation, the topology of the brain is preserved and this prevents folding, which is physically impossible (Tom Vercauteren, Xavier Pennec, & Ayaches, 2009). According to the diffeomorphic mapping, the transformation $\varphi : \Omega \to \Omega$, which maps a point on the reference image $x \in \Omega$ to another point on the source image, is obtained by integrating the small time-dependent vector field $v_t : \Omega \to R^3, t \in [0, 1]$ starting from the identity transformation ϕ_0 at time $t = 0$ to the final one at time $t = 1$,

$$\varphi = \phi_1 = \phi_0 + \int_0^1 v_t(\phi_t)\mathrm{d}t. \tag{3}$$

The optimal vector field at each time step is obtained by solving the optimization problem:

$$\hat{v} = \arg\min_{v} \left(S(I \circ \varphi^{-1}, J) + \int_0^1 \|Lv_t\|_{L^2}^2 \mathrm{d}t \right). \tag{4}$$

The objective function, which consists of the similarity measure S between the two images and a regularization term to control the smoothness of the vector field, prevents the obtained transformation from violating the topological correspondence. In SyN, we simultaneously optimize two sets of the optimal vector fields that transform the reference image and the source image to intermediate images between them. The similarity measure is a cross correlation between the two deformed images on the space of the intermediate image.

Since the obtained segmentation often preserves the topology of brain structures, the diffeomorphic transformation-based registration does not produce blob-like errors. On the other hand, the transformation is prone to errors near the boundaries between different regions, because the smoothness prior attached as the regularization term prefers rather simple correspondence; hence, it places less emphasis on the precise alignment of the two images, introducing some discrepancy near the regional boundaries.

2.2. Deep neural network

Deep neural networks (DNNs) have made significant progress and have been used in many machine-learning applications. Convolutional neural networks (CNNs), which are variants of DNNs, have shown excellent performance in a variety of image processing applications. Each CNN comprises a number of convolutional and pooling layers; the former convolves the input image with a convolutional kernel and the latter introduces a certain shift-invariance. The convolutional kernels are optimized to perform optimally in a specific supervised learning task, based on a training dataset that consists of pairs containing an input image and output label. This optimization of the convolutional operations enabled CNNs to exhibit the best performance in a data-driven fashion on various image processing tasks, such as image recognition and semantic segmentation (Krizhevsky, Sutskever, & Hinton, 2012; Long, Shelhamer, & Darrell, 2015; Ren, He, Girshick, & Sun, 2015).

Inspired by the success of DNNs, methods based on deep learning have been applied to image segmentation problems of brains (Litjens et al., 2017). Many of these methods were in 2D, i.e., they were applied to 2D slices of brain images, which cannot make use of 3D information of the brain structure (de Brébisson & Montana, 2015; Zhang et al., 2015). VoxResNet (Chen et al., 2017) achieved the best performance among various DNN-based methods on the 2013 MICCAI MRBrainS challenge (Mendrik, Vincken, Kuijf, Breeuwer, Bouvy, De Bresser, Alansary, De Bruijne, Carass, El-Baz, et al., 2015). This method employs 3D convolutions, such that it directly outputs a 3D label map $L_I \in \{0, 1\}^{L\times W\times H\times C}$ given a 3D brain image $I \in \mathcal{R}^{L\times W\times H\times D}$ as input, where L, W, H are the number of the voxels of length, width, and height of the image, respectively. Owing to its residual structures (He, Zhang, Ren, & Sun, 2016), VoxResNet makes it possible to avoid the gradient vanishing problem, thereby enabling the deepest 3D convolutional architecture to be trained thus far.

In a recent paper it was pointed out that registration-based methods and DNN-based methods are complimentary in terms of segmentation error (Pai, Teng, Blair, Kallenberg, Dam, Sommer, Igel, & Nielsen, 2017). Registration-based methods are advantageous because of their preservation of the topology of brain regions, but are disadvantageous because they sometimes introduce segmentation errors near the boundaries between different regions. These methods employ a strong prior information for the spatial transformation to avoid foldings, which may work against the exact alignment of two images. Segmentation based on DNNs, on the other hand, gives relatively precise prediction near the regional boundaries, leading to good segmentation, but is instead sensitive to image noise in uniform regions, causing blob-like errors.

3. Method

Although DNNs have shown good performance in terms of brain image segmentation, they require a number of meticulously annotated brain images for their training. A DNN trained on a specific dataset often performs poorly when segmenting images acquired by using different imaging experimental settings or from different species. Since annotating 3D brain images requires laborious works of expert anatomists, it is often difficult to prepare a dataset to train the DNN for a specific kind of images for which segmentation is required. Thus, there is considerable demand for methods capable of facilitating image annotation.

A variety of data augmentation techniques are available to increase the effective size of the training dataset. These techniques were mostly developed for use in the field of computer vision: cropping, flipping, scaling, applying spatial transformations, etc. Preparation of a much larger dataset would be more effective than data augmentation, because the latter approach essentially involves usage of the same data a second time. However, the former approach may not be feasible.

In this study, we propose a semi-supervised training algorithm that trains a DNN-based segmentation model based not only on annotated (labeled) brain images but also on unlabeled brain images. When considering real-world applications, we often have a few annotated brain images and a large number of unlabeled brain images. Based on the naïve idea, we simply attach pseudo-labels to the unlabeled brain images. This enables us to train a DNN-based segmentation model on the integrated dataset consisting of originally labeled brain images and originally unlabeled but pseudo-labeled brain images. We use label propagation (Avants et al., 2008) to assign pseudo-labels to the unlabeled brain images; segmentation itself is performed by transferring the labels attached to the annotated brain images.

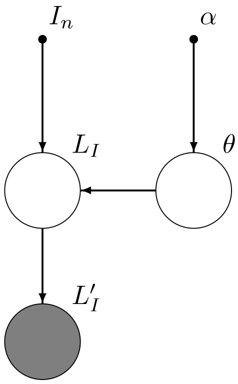

Fig. 1. Graphical representation of our proposed method for a single unlabeled brain image I. L_I and L'_I are unobservable true label and observable pseudo-label images, respectively. θ and α are parameters of the DNN and its hyperparameter, respectively.

The disadvantage of the above-mentioned idea, however, is that the model is also trained on pseudo-labeled images of which the labels inherently include mislabels produced by the label propagation, especially near the regional boundaries. Our solution to this problem is to propose a probabilistic model to incorporate these mislabels attached by the label propagation; the estimation of the probabilistic model corresponds to the identification of the process whereby these mislabels are generated. More concretely, in our probabilistic model, a DNN f parameterized with θ gives the probability distribution for a true label image L_I, given an input brain image I; we assume that it follows a categorical (multi-nomial) distribution:

$$p(L_I|I, \theta) = \mathrm{Cat}(L_I|f(I, \theta)). \tag{5}$$

A pseudo-labeled image L'_I is produced by adding spatial noise to the true labeled image L_I subsequently. Hence, the joint distribution is given by

$$p(L'_I, L_I, \theta|I, \alpha) = p(L'_I|L_I)p(L_I|I, \theta)p(\theta|\alpha), \tag{6}$$

of which the graphical model (Bishop, 2006) is shown in Fig. 1.

The likelihood function $p(L'_I|L_I)$ of true label image L_I given pseudo-label L'_I is

$$p(L'_I|L_I) = \mathrm{softmax}(D(L'_I)/\sigma), \tag{7}$$

where $D(L'_I)$ denotes distance transform (Borgefors, 1986) and σ is the parameter for controlling the output's randomness. Although there are various choices of label smoothing such as a Gaussian filter (Schindler, 2012), we used the distance transform to deal with the boundaries of brain areas attached by different labels in an effective manner. Let $D(L'_I)(\mathbf{x})$ represent the vector at voxel $\mathbf{x}$ after applying the distance transform to the pseudo-label L'_I; the dimensionality of this vector is the same as the number of categories. The cth component of the vector is given by

$$D(L'_I)(\mathbf{x})_c = \begin{cases} |\mathbf{x} - \mathbf{b}_c| & (L'_I(\mathbf{x})_c = 1) \\ 1 - |\mathbf{x} - \mathbf{f}_c| & (\text{otherwise}), \end{cases} \tag{8}$$

where $\mathbf{b}_c$ is the closest voxel to $\mathbf{x}$ in terms of the Euclidean distance that satisfies $L'_I(\mathbf{b}_c)_c = 0$, and $\mathbf{f}_c$ is the closest voxel to $\mathbf{x}$ that satisfies $L'_I(\mathbf{f}_c)_c = 1$. The softmax function is applied to ensure that the values are in [0, 1] such that this observation process is valid as a probability distribution. Further, $p(\theta|\alpha)$ is the prior probability of parameter θ, which is parameterized by hyperparameter α.

Because the true label image is a hidden variable in our probabilistic model, we rely on the expectation–maximization (EM) algorithm (Bishop, 2006) to perform maximum a posteriori (MAP) estimation of parameter θ, given the hyperparameter α. The log posterior distribution of parameter θ given an observable pseudo-label L'_I and an assumed hyperparameter α is

$$\ln p(\theta|L'_I, I, \alpha) = \ln p(L'_I|I, \theta) + \ln p(\theta|\alpha) + const. \tag{9}$$

Let $q(L_I)$ be an arbitrary probability distribution (termed a trial distribution) of true label L_I. Then, we have

$$\begin{aligned}
\ln p(L'_I|I, \theta) &= \int q(L_I) \ln p(L'_I|I, \theta) \mathrm{d}L_I && (10)\\
&= \int q(L_I) \ln \frac{p(L'_I, L_I|I, \theta)}{p(L_I|L'_I, I, \theta)} \mathrm{d}L_I && (11)\\
&= \int q(L_I) \ln p(L'_I, L_I|I, \theta) \mathrm{d}L_I && (12)\\
&\quad - \int q(L_I) \ln q(L_I) \mathrm{d}L_I \\
&\quad + \mathrm{KL}(q(L_I) \parallel p(L_I|L'_I, I, \theta)) \\
&\geq \int q(L_I)\{\ln p(L'_I|L_I) + \ln p(L_I|I, \theta)\} \mathrm{d}L_I \\
&\quad - \int q(L_I) \ln q(L_I) \mathrm{d}L_I. && (13)
\end{aligned}$$

Here, $\mathrm{KL}(q(L_I) \parallel p(L_I|L'_I, I, \theta))$ is the Kullback–Leibler divergence between $p(L_I|L'_I, I, \theta)$ and $q(L_I)$. Thus, the log posterior distribution, eq. (9), is lower bounded by

$$\ln p(\theta|L'_I, I, \alpha) \geq \int q(L_I) \ln p(L_I|I, \theta) \mathrm{d}L_I + \ln p(\theta|\alpha) + const., \tag{14}$$

which is the free energy function. The EM algorithm repeats the E-step that makes the trial distribution becomes

$$q(L_I) = p(L_I|L'_I, I, \theta) \propto p(L'_I|L_I) p(L_I|I, \theta), \tag{15}$$

and the M-step that maximizes the free energy function with respect to parameter θ.

We reduced the computational cost by adopting the hard-EM approximation (Papandreou, Chen, Murphy, & Yuille, 2015). In the E-step, we obtain the MAP estimate of the true label:

$$\hat{L}_I = \arg\max_{L_I} p(L_I|L'_I, I, \theta), \tag{16}$$

and in the M-step, we maximize the free energy function into which the MAP estimate is plugged, with respect to the parameter θ:

$$\begin{aligned}
&\int q(L_I) \ln p(L'_I, L_I|I, \theta) \mathrm{d}L_I + \ln p(\theta|\alpha) + const. && (17)\\
&\approx \ln p(L'_I, \hat{L}_I|I, \theta) + \ln p(\theta|\alpha) + const. && (18)\\
&= \ln p(\hat{L}_I|I, \theta) + \ln p(\theta|\alpha) + const. && (19)
\end{aligned}$$

That is, we replaced the trial distribution, eq. (15), with a delta distribution centered at the MAP estimate.

Accordingly, our proposed semi-supervised learning algorithm is shown in Algorithm 1. Since training a DNN model is time consuming, we can terminate our algorithm at step 2 (that is, without EM), which is termed the simple version of our proposed method.

The algorithm presented here assumes that only one atlas is provided. We introduce an ensemble learning approach for enabling to use multiple atlases. Suppose that we have M atlases with brain images and N unlabeled brain images. To allow our evaluation to be rather independent from the possible bias to each atlas, we prepare M datasets each consisting of one atlas with brain image and $M + N - 1$ unlabeled images, where each of the sole atlas is different between different datasets. We independently apply the proposed method to the M datasets, and estimate M parameters $(\theta_1, \ldots, \theta_M)$. The average probability distribution for the true label given an input image over the M datasets is obtained as $(1/M)\sum_{i=1}^{M} p(L_I|I, \theta_i)$. Similarly, we can estimate the label probability distribution for any number of atlases less than or equal to M.

Algorithm 1: Semi-supervised learning method.

```
INPUT: atlas $\{I_0, L_{I_0}\}$, brain images $\{I_1, I_2, \cdots, I_N\}$
1: Assign pseudo-labels $\{L'_{I_1}, \cdots, L'_{I_N}\}$ based on image registration using the atlas.
2: Estimate initial CNN parameters $\theta$ trained on the atlas and on the pseudo-atlases. The proposed method (simple) is finished at this step.
3: $K = 1$
while $K \leqq 5$ do
    4: E-step, compute MAP estimate of the true label for each unlabeled brain image,
       $\hat{L}_{I_n} = \operatorname{argmax}_{L_{I_n}} p(L_{I_n}|L'_I, I, \theta).$
    5: M-step, update CNN parameters given the estimated true label images,
       $\theta \leftarrow \operatorname{argmax}_{\theta'}\{\ln p(L_{I_0}|I_0, \theta') + \sum_{n=1}^{N} \ln p(\hat{L}_{I_n}|I_n, \theta') + \ln p(\theta'|\alpha)\}.$
    6: $K = K + 1$
end while
OUTPUT: $\theta$
```

4. Experimental evaluation

4.1. Evaluation metric

We compared our proposed method with the baseline methods by two metrics: the Dice coefficient (DC) and absolute volume difference (AVD). Let G_a be a set of voxels annotated by experts as a certain region, i.e., with a single label a, and S_a be another set of voxels to which a segmentation method assigned the same label a. The DC of this region a is

$$\mathrm{DC}(G_a, S_a) = 2\frac{|G_a \cap S_a|}{|G_a| + |S_a|}, \tag{20}$$

where $|\cdot|$ is the number of voxels in the set. The DC measures the similarity of the two sets, which is equivalent to the F-measure, that is, the harmonic mean of precision and recall. The AVD of a single region a is

$$\mathrm{AVD}(G_a, S_a) = \frac{|V_{g_a} - V_{s_a}|}{V_{g_a}}, \tag{21}$$

where V_{g_a} is the volume assigned by the label a of this region as the ground truth (human annotation) and V_{s_a} is that of the segmentation result. Further, $|\cdot|$ denotes the absolute value in the above equation. The AVD is a popular measure, because the volumes of brain regions can be important biomarkers for clinical use.

Table 1
Dice coefficient (DC) for each of cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM). Each value in mean ± std. The highest mean DC using only one atlas is in bold text. #labeled means the number of atlases (labeled images) and #unlabeled is the number of unlabeled images.

Method	#labeled	#unlabeled	CSF	GM	WM	Average
LP (SyN)	1	0	0.78 ± 0.06	0.82 ± 0.03	0.76 ± 0.02	0.79 ± 0.02
VoxResNet	1	0	0.61 ± 0.17	0.87 ± 0.06	0.87 ± 0.05	0.78 ± 0.07
VoxResNet + data augmentation	1	0	0.71 ± 0.15	0.88 ± 0.06	0.88 ± 0.03	0.82 ± 0.07
Proposed (simple)	1	9	0.81 ± 0.05	0.89 ± 0.02	0.87 ± 0.02	$\mathbf{0.86} \pm 0.02$
Proposed (full)	1	9	$\mathbf{0.82} \pm 0.05$	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.88} \pm 0.02$	$\mathbf{0.86} \pm 0.02$
VoxResNet	10	0	0.87 ± 0.04	0.93 ± 0.01	0.92 ± 0.02	0.91 ± 0.02

Table 2
Absolute volume difference (AVD) for each of cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM). Each value in mean ± std., the lowest mean AVD using only one atlas is in bold text. #labeled means the number of atlases (labeled images) and #unlabeled is the number of unlabeled images.

Method	#labeled	#unlabeled	CSF	GM	WM	Average
LP (SyN)	1	0	0.18 ± 0.13	0.06 ± 0.05	0.08 ± 0.06	0.11 ± 0.06
VoxResNet	1	0	0.65 ± 0.89	0.14 ± 0.11	0.18 ± 0.12	0.32 ± 0.34
VoxResNet + data augmentation	1	0	0.45 ± 0.59	0.12 ± 0.11	0.12 ± 0.11	0.23 ± 0.25
Proposed (simple)	1	9	0.15 ± 0.10	0.05 ± 0.03	$\mathbf{0.07} \pm 0.05$	0.09 ± 0.04
Proposed (full)	1	9	$\mathbf{0.11} \pm 0.09$	$\mathbf{0.04} \pm 0.04$	0.09 ± 0.06	$\mathbf{0.08} \pm 0.04$
VoxResNet	10	0	0.09 ± 0.07	0.05 ± 0.02	0.06 ± 0.04	0.07 ± 0.03

4.2. Human brain image

In this experiment we evaluated our proposed method and the baseline methods, by using human MR images registered at the Internet Brain Segmentation Repository (IBSR). This dataset contains T1-weighted brain images ($K = 1$) of 18 subjects with a 1.5-mm slice thickness ($256 \times 128 \times 256$). Expert anatomists annotated each voxel of the scans as belonging to one of four regions: background, cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM). We first aligned their spatial resolution to be isotropic (1 mm × 1 mm × 1 mm) using the bilinear interpolation of the images. Subsequently, we pre-processed the T1-weighted brain images following the paper in which VoxResNet was originally proposed (Chen et al., 2017), where we used a difference of Gaussian filter and applied the contrast limited adaptive histogram equalization (CLAHE) to enhance edges in the images (Pizer, Amburn, Austin, Cromartie, Geselowitz, Greer, ter Haar Romeny, Zimmerman, & Zuiderveld, 1987).

Among 18 subjects, 6 were used for the final test and 2 subjects were used for validation, because of the avoidance of possible bias due to the usage of only 1 subject. We randomly chose these validation and test subjects. Using the remaining 10 subjects, we constructed six segmentation methods with different settings.

- image registration-based method with one atlas. (i.e., label propagation of SyN implemented using ANTs (Avants, Tustison, Song, Cook, Klein, & Gee, 2011), http://stnava.github.io/ANTs/)
- VoxResNet trained on one atlas
- VoxResNet trained on one atlas with data augmentation applying spatial transformations on both the brain image and label image of the one atlas.
- VoxResNet trained by the simple version of our proposed method employing one atlas and nine unlabeled images
- VoxResNet trained by the full version of our proposed method employing one atlas and nine unlabeled images
- VoxResNet trained on the set of ten atlases

Here, an atlas refers to a pair of T1-weighted brain image from a single subject and the corresponding label image. It should be noted that step 1 in our algorithm (see Algorithm 1) is the image registration. For the sake of fair comparison, we employed the estimated transformations in step 1 of our algorithm as the spatial transformations applied in the (VoxResNet + data augmentation, one atlas) setting. Since there are several hyperparameters in SyN and VoxResNet, they were tuned to obtain the best performance in terms of DC for two validation images. In the proposed methods, we set the parameter σ in the softmax function at 1 mm, which is the same as the voxel size of the image. Repeating the experiments five times with different combinations of the training, validation, and test brain images, we evaluated the mean and standard deviation (std.) of the two metrics, DC and AVD, for each of the six settings above.

Tables 1 and 2 list the mean and std. for testing DC and AVD, respectively. Under the condition that one atlas is available, our proposed method (full version) exhibited the best performance in terms of both DC and AVD, which is almost comparable to the performance of VoxResNet using 10 atlases. A comparison of SyN and VoxResNet shows that the DNN (in this case, VoxResNet) was not effective when only one labeled image was employed. In this case the registration-based method (SyN) was fairly effective. In our method, semi-supervised learning employing the registration method was effective in leveraging unlabeled brain images by probabilistically attaching pseudo-labels based on the image registration.

Fig. 2.A shows the segmented brain images of two subjects, in which green, yellow, and brown represent CSF, gray matter, and white matter, respectively. We choose the two subjects: the upper one (subject 1) is of the typical AVD value by the proposed method and the bottom one (subject 2) is of the worst AVD value by the proposed method. For these two subjects, VoxResNet trained by one atlas (Fig. 2.A) assigned the white matter label to a wider region of voxels than the ground truth, especially in subject 2. On the other hand, VoxResNet trained by the simple and full versions of our method assigned the same label less than the ground truth. Fig. 2.B shows the image of the segmentation errors (red region) between the manual segmentation and other segmentations. This result shows that the mislabels tend to be located at the white and gray matter boundaries, and the error region of the label propagation (by SyN) is larger than that of the proposed method.

Fig. 2.C visualizes the mean and std. of the AVD and DC values (Tables 1 and 2). We applied a paired t-test to the test performance, so that the asterisk indicates the significant difference ($p < 0.05$) between the proposed method (full version) and the other methods. We found that the VoxResNet sometimes mislabeled the white and gray matters (Fig. 2.A). Correspondingly, Fig. 2.C shows the significant difference in AVD between the VoxResNet and the proposed method (full version). Similarly,

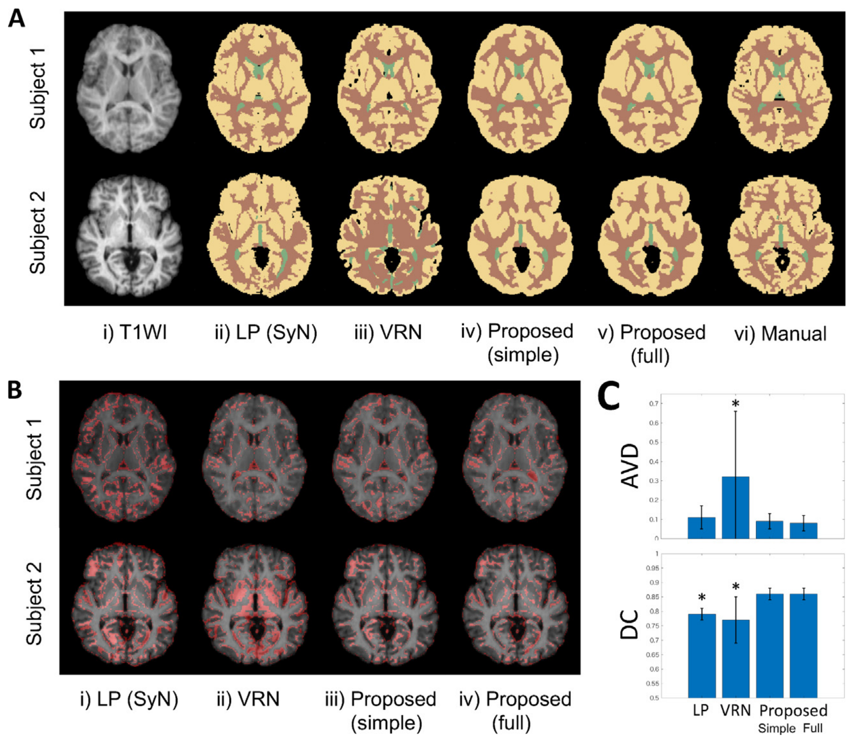

Fig. 2. Segmentation result in IBSR. (A) The segmented labels of the two subjects. The green, yellow, and brown represent CSF, gray matter, and white matter, respectively. (i) Original T1-weighted image; (ii) segmentation result of label propagation (LP) based on one atlas; (iii) segmentation result of VoxResNet(VRN) trained based on one atlas; (iv) segmentation result of VoxResNet trained based on one atlas and nine unlabeled images by the simple version; (v) segmentation result of VoxResNet trained based on one atlas and nine unlabeled images by the full version of our proposed method; (vi) ground truth label. (B) The images of the segmentation errors. The red region indicates the mislabeling between the methods' and manual labels. (i) Segmentation errors of label propagation (LP) based on one atlas; (ii) segmentation errors of VoxResNet trained based on one atlas; (iii) segmentation errors of VoxResNet trained based on one atlas and nine unlabeled images by the simple version; (iv) segmentation errors of VoxResNet trained based on one atlas and nine unlabeled images by the full version of our proposed method. (C) The bar plot of the average performance of mean and std. of AVD and DC. The upper and bottom panels show the AVD and DC, respectively. An asterisk means the significant difference between the proposed method (full version) and the other methods with the paired t-test ($p < 0.05$) . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

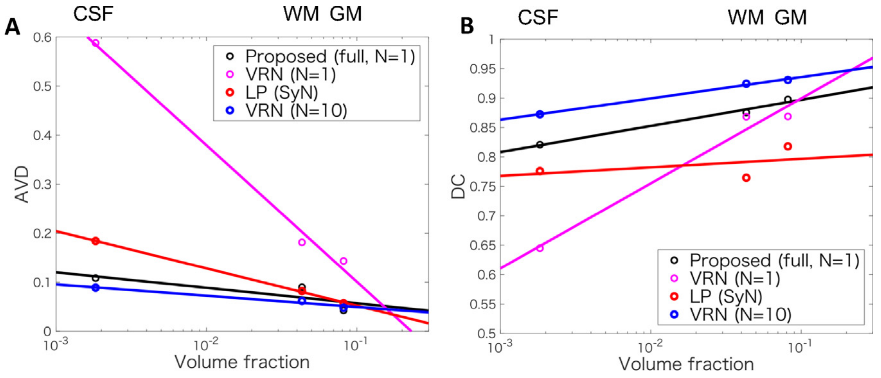

Fig. 3. How AVD (left) and DC (right) are dependent on the volume fraction of brain regions in IBSR. The horizontal axis is the logarithm of the volume fraction of brain region: CSF, WM and GM. The vertical axis is the AVD (left) and DC (right). The black, blue, red and magenta circles represent the means of the AVD and DC of the proposed method (full version), VoxResNet(VRN) with 10 atlases, label propagation (LP) with SyN and VoxResNet with one atlas, respectively. The colored lines denote the linear regression of the values of AVD and DC for the test images . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

observation in Fig. 2.B suggested that the errors on the white and gray matter boundaries led to the significant difference between the proposed method (full version) and the other methods. That is, the proposed method successfully reduced the bias so that its segmented image is much closer to the ground truth than those by the other methods.

The average computational time of the proposed method (full version) was about 65 h with 5 iterations in our EM algorithm, and the average computational time of the simple VoxResNet in our implementation was about 12 h, when both were working on a cluster machine with a CPU of Intel Xeon E5-2640 v4 2.40 GHz and 4 GPUs of Nvidia GeForce 1080Ti.

Fig. 3 shows the relation between the metrics and the volume fraction of the brain regions. The horizontal axis is the logarithm of the volume fraction of the brain region without background region. The black, blue, red and magenta circles represent the means of the AVD and DC of the proposed method (full version), VoxResNet with 10 atlases, label propagation with SyN and VoxResNet with one atlas, respectively. The lines with the corresponding color show the linear regression of the AVD and

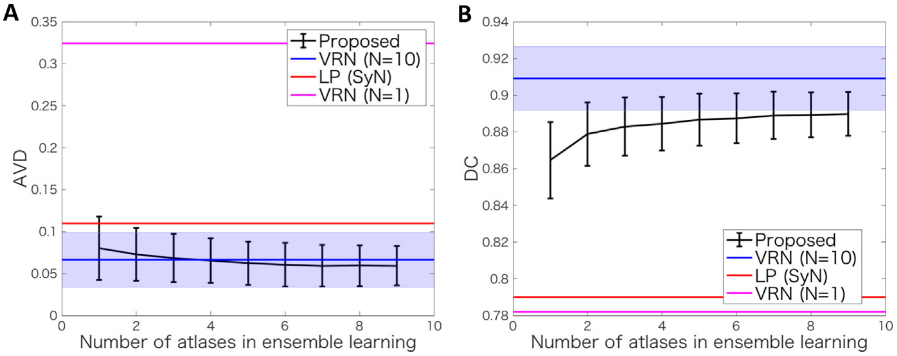

Fig. 4. AVD (left) and DC (right) when the number of atlases was changed in the proposed method (ensemble). The mean performances of the proposed method (ensemble), VoxResNet(VRN) using 10 atlases, label propagation (LP) using SyN and VoxResNet using one atlas are represented by black, blue, red and magenta colors, respectively. The horizontal axis indicates the number of atlases used in the proposed method (ensemble). We show the std. of the AVD of the proposed method (ensemble) with the black errorbar and that of VoxResNet using 10 atlases with blue region. The std. of the other methods was omitted for visibility . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

DC for the test images. The AVD decreases and the DC increases when the volume fraction increases in all cases. The regression slopes of the proposed method and the VoxResNet with 10 atlases are similar, whereas the slope of the VoxResNet with one atlas is very different from the others. This is because the AVD and DC of CSF is rather high and low, respectively, in the VoxResNet with one atlas. This result suggests that the VoxResNet with one atlas would not be reliable, especially in the small region like CSF. Comparing to the VoxResNet with one atlas, the proposed method with one atlas worked well even in CSF.

So far, we showed the results by the proposed method when it employed one atlas. Here, we generalize the proposed method into usage of multiple atlases (see Method), which we call the proposed method (ensemble). Fig. 4 shows the AVD and DC of the proposed method (ensemble) compared with the VoxResNet with 10 atlases, the label propagation with SyN and the VoxResNet with one atlas, which are represented by the black, blue, red and magenta curves or lines, respectively. The horizontal axis indicates the number of atlases, which was used in the proposed method (ensemble). The AVD (DC) with respect to the number of atlases monotonically decreased (increased). This figure shows that the proposed method (ensemble) using multiple atlases outperformed the label propagation and the VoxResNet trained on one atlas both in term of AVD and DC. The AVD of the proposed method was lower than the VoxResNet with 10 atlases when the number of atlases in the ensemble learning was more than 4. The DC of the proposed method (ensemble) was lower than the VoxResNet with 10 atlases, although the distribution of DC of the proposed method (shown by the black errorbar) was close to that of the VoxResNet with 10 atlases (shown by the blue region). This result also suggests that the proposed method (ensemble) with 4 atlases outperformed any other method, including the VoxResNet with 10 atlases in terms of AVD.

4.3. *Marmoset brain image*

Using our original dataset consisting of marmoset brain images, we also compared our segmentation method with the baseline methods. Okano and colleagues at the Riken Center for Brain Science have imaged marmoset *ex vivo* brains by diffusion-weighted magnetic resonance imaging (DWI) using a b-value of 5,000, employing a Bruker 9.4 T scanner (Okano et al., 2015, 2016). They provided us with 13 marmoset DWI ($b0$) images, all of which were annotated by expert anatomists. The task here is to reproduce these human annotations; that is, to achieve voxel-wise segmentation into six classes: background, white matter (WM), right cerebral cortex (RCC), left cerebral cortex (LCC), subcortical gray matter (SGM), and cerebellum cortex (CC). We pre-processed the DWI images using spherical harmonics ($K = 128$) following the previous study (Schnell et al., 2009), and then standardized the image intensities such that the mean and variance for each image are zero and unity, respectively.

We divided the 13 images from different animals into 2 validation images to avoid possible bias due to the usage of only 1 animal, 6 test images, and the remaining training images. The number of test images, six, was set to the same as the number of test subjects in the application to human MR images. We randomly chose these validation and test images. Using five images from the remainder, we compared five segmentation methods with different settings: label propagation of SyN with one atlas, VoxResNet trained based on one atlas, VoxResNet trained based on one atlas and four unlabeled DW images by the simple version, VoxResNet trained based on one atlas and four unlabeled DW images by the full version of our proposed method, and VoxResNet trained based on five atlases. Both SyN and VoxResNet include hyperparameters that are tuned as to maximize DC for the two validation images. In the proposed methods, we set the parameter σ in the softmax function at 1 mm. This is larger than the voxel size of the image, but the preliminary analysis found that setting the parameter as the same as the voxel size 0.3 mm resulted in poorer performance than that with the current setting.

Tables 3 and 4 provide the mean and std. of DC and AVD, respectively. Similar to the experiment on the human MRI dataset, both the simple and full versions of our proposed method presented results that are more accurate and stable than those by the baseline methods, when using the same number of labeled images (atlases). Moreover, the performance of our full version is comparable to that of VoxResNet trained on five atlases. This means our semi-supervised learning method successfully performed the segmentation that is generalizable into different individuals, by appropriately incorporating the similarity in DW images between different individuals.

Fig. 5.A shows the segmented brain images of two subjects, in which green, yellow, and brown represent CSF, gray matter, and white matter, respectively. We choose the two subjects: the upper one (subject 1) is of the typical AVD value by the proposed method and the bottom one (subject 2) is of the worst AVD value by the proposed method. In subject 1, both of the VoxResNet and the proposed method (simple version) produced mislabeling in the gray matter of the occipital cortex (bottom region). Because the signal of the bottom region in subject 2 seemed to have been attenuated, the segmentation by the VoxResNet was directly affected by this attenuation. On the other hand, the segmentation by the proposed method (full version) covered wider region of the

Table 3
Dice coefficient (DC) for each of white matter (WM), right cerebral cortex (RCC), left cerebral cortex (LCC), subcortical gray matter (SGM), and cerebellum cortex (CC). Each value in mean ± std. The highest mean DC using only one atlas is in bold text. #labeled means the number of atlases (labeled images) and #unlabeled is the number of unlabeled images.

Method	#labeled	#unlabeled	WM	RCC	LCC	SGM	CC	Average
LP (SyN)	1	0	0.73 ± 0.01	0.83 ± 0.02	0.81 ± 0.02	0.85 ± 0.02	0.73 ± 0.05	0.79 ± 0.02
VoxResNet	1	0	0.88 ± 0.02	0.93 ± 0.02	0.93 ± 0.02	0.89 ± 0.04	0.78 ± 0.14	0.88 ± 0.05
Proposed (simple)	1	4	**0.91** ± 0.01	0.95 ± 0.01	**0.95** ± 0.01	0.93 ± 0.00	0.92 ± 0.01	0.93 ± 0.01
Proposed (full)	1	4	**0.91** ± 0.01	**0.96** ± 0.01	**0.95** ± 0.01	**0.94** ± 0.00	**0.93** ± 0.01	**0.94** ± 0.01
VoxResNet	5	0	0.92 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.95 ± 0.00	0.93 ± 0.01	0.94 ± 0.01

Table 4
Absolute volume difference (AVD) for each of white matter (WM), right cerebral cortex (RCC), left cerebral cortex (LCC), subcortical gray matter (SGM), and cerebellum cortex (CC). Each value in mean±std. The lowest mean AVD using only one atlas is in bold text. #labeled means the number of atlases (labeled images) and #unlabeled is the number of unlabeled images.

Method	#labeled	#unlabeled	WM	RCC	LCC	SGM	CC	Average
LP (SyN)	1	0	0.049 ± 0.017	0.037 ± 0.035	0.052 ± 0.025	0.052 ± 0.043	0.091 ± 0.022	0.056 ± 0.017
VoxResNet	1	0	0.038 ± 0.023	0.022 ± 0.029	0.039 ± 0.017	0.050 ± 0.012	0.221 ± 0.210	0.074 ± 0.051
Proposed (simple)	1	4	0.024 ± 0.020	0.016 ± 0.016	**0.016** ± 0.024	0.044 ± 0.013	0.036 ± 0.019	0.027 ± 0.013
Proposed (full)	1	4	**0.023** ± 0.016	**0.014** ± 0.013	0.021 ± 0.011	**0.010** ± 0.007	**0.034** ± 0.018	**0.020** ± 0.005
VoxResNet	5	0	0.042 ± 0.014	0.012 ± 0.016	0.026 ± 0.009	0.009 ± 0.005	0.035 ± 0.025	0.025 ± 0.005

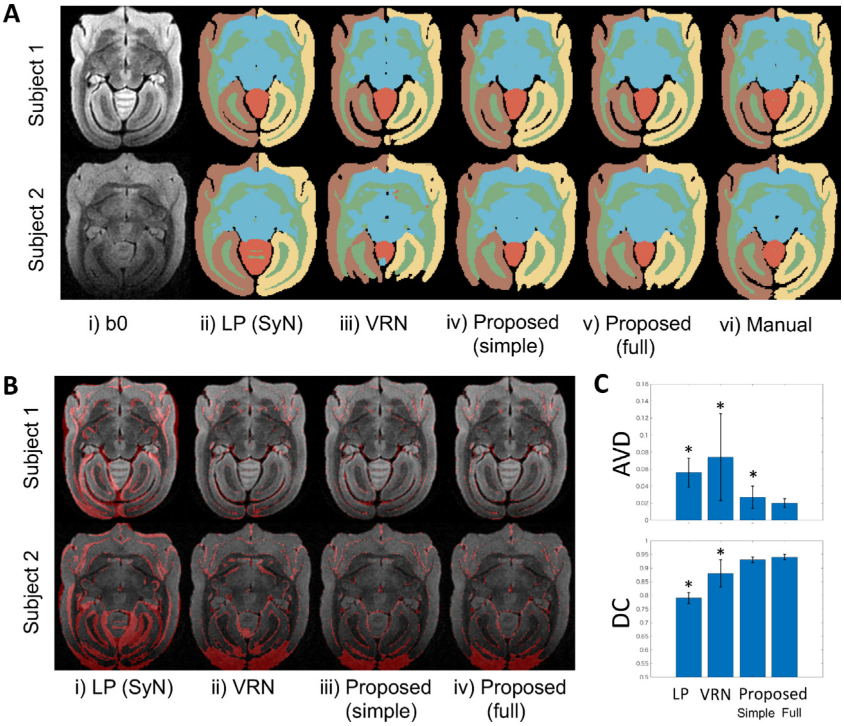

Fig. 5. Segmentation results in the marmoset datasets. (A) The segmented labels of the two subjects. The green, yellow, and brown represent CSF, gray matter, and white matter, respectively. (i) Original T1-weighted image; (ii) segmentation result of label propagation (LP) based on one atlas; (iii) segmentation result of VoxResNet(VRN) trained based on one atlas; (iv) segmentation result of VoxResNet trained based on one atlas and nine unlabeled images by the simple version; (v) segmentation result of VoxResNet trained based on one atlas and nine unlabeled images by the full version of our proposed method; (vi) ground truth label. (B) The images of the segmentation errors. The red region indicates the mislabeling between the methods' and manual labels. (i) Segmentation errors of label propagation (LP) based on one atlas; (ii) segmentation errors of VoxResNet trained based on one atlas; (iii) segmentation errors of VoxResNet trained based on one atlas and nine unlabeled images by the simple version; (iv) segmentation errors of VoxResNet trained based on one atlas and nine unlabeled images by the full version of our proposed method. (C) The bar plot of the mean and std. of AVD (left) and DC (right). An asterisk means the significant difference between the proposed method (full version) and the other methods with the paired t-test ($p < 0.05$) . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

occipital cortex than by the VoxResNet. Fig. 5.B shows difference (red region) between the manual and each method. Similar to the case of the IBSR dataset, the mislabels mostly exist in the white and gray matter boundaries, and moreover, the error regions of the label propagation by SyN are likely larger than those of the proposed method.

Fig. 5.C depicts the mean and std. of the AVD and DC values (Tables 3 and 4). An asterisk in these panels signifies the significant difference (paired t-test, $p < 0.05$) of AVD or DC between the proposed method (full version) and the baseline method. Fig. 5.A shows the difference in labeling in the gray matter of the VoxResNet and the proposed method (full version), suggesting that the VoxResNet introduced mislabels to the gray matter regions. Due to such mislabels, there has been the significant difference between the proposed method (full version) and the other method (Fig. 5.C). Similarly, observation in

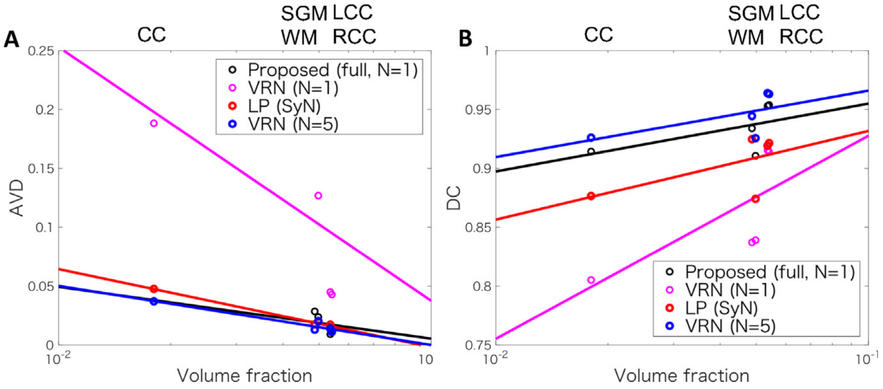

Fig. 6. How AVD (left) and DC (right) are dependent on the volume fraction of brain regions in the marmoset dataset. The horizontal axis is the logarithm of the volume fraction of brain region: CC, SGM, WM, LCC and RCC. The black, blue, red and magenta circles represent the means of the AVD (DC) of the proposed method (full version), VoxResNet(VRN) with 5 atlases, label propagation (LP) with SyN and VoxResNet with one atlas, respectively. The colored lines denote the linear regression of the values of AVD and DC for the test images . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 5.B suggested that the errors on the white and gray matter boundaries led to the significant difference between the proposed method (full version) and the label propagation. These indicate that the proposed method (full version) reduced this bias, so that its segmented image is much closer to the ground truth than those by the other methods in the marmoset datasets.

Here, we show the relation between the metrics and the volume fraction of the brain regions using the marmoset dataset in Fig. 6. The horizontal axis is the logarithm of the volume fraction of the brain region without the background region. The black, blue, red and magenta circles represent the means of the AVD and DC of the proposed method (full version), VoxResNet with 10 atlases, label propagation with SyN and VoxResNet with one atlas, respectively. The lines with the corresponding color show the linear regression of AVD and DC for the test images. The ascending order of the volume fraction is as follows: CC, SGM, WM, LCC and RCC. This result was similar to that of the human dataset (Fig. 3). The AVD decreased (the DC increased) when the volume fraction increased in all the methods. The regression slopes of the proposed method, the label propagation and the VoxResNet with 10 atlases behaved similarly, whereas the slope of the VoxResNet with one atlas was very different from the others. This observation also suggests that the VoxResNet with one atlas would not be reliable, especially in the small region. Comparing to the VoxResNet with one atlas, the proposed method with one atlas worked well in the small region, such as cerebellum cortex (CC) in the marmoset images.

5. Discussion

Our semi-supervised image segmentation method achieved better segmentation than the existing registration-based and DNN-based methods, given the same number of labeled images; the advantage of our method was prominent especially on small regions such as the CSF in the human brain image and the cerebellum cortex in the marmoset brain image. Because the loss function for training a DNN, eq. (19), is the sum of voxel-wise losses, the usual supervised learning of the DNN puts larger emphasis on larger regions. Our semi-supervised learning method, on the other hand, successfully segmented these small regions even with limited amount of annotation, owing to introducing unlabeled images into the training process.

When a DNN was trained based only on one labeled image, the segmentation accuracy of the trained DNN varied as to reflect the characters of the individual image used for training. The data augmentation technique, which essentially entails the secondary use of the same image, cannot enough reduce such dependence of the segmentation accuracy. On the other hand, our proposed method succeeded in performing reasonably good segmentation in a stable manner, even when a single labeled image is provided for training; that is, leveraging unlabeled images for training successfully regularized the trained DNN.

There is an existing study that employed the EM algorithm in the scenario of weakly semi-supervised learning for image segmentation problems (Papandreou et al., 2015). Our method is different from this previous method, because our method incorporates the stochastic process of pseudo-labeling of MRI images as a particular probabilistic model, whereas the previous study above regarded image-level labels or bounding-box annotations as observed variables and the pixel-level segmentation as a hidden variable, to deal with weakly semi-supervised learning situations for general 2D images. Another study using the EM algorithm presented a semi-supervised clustering approach for brain segmentation (Portela, Cavalcanti, & Ren, 2014). It was necessary to perform clustering for segmenting new images, according to this method. Our method based on DNNs has an advantage in the computation cost when applied to new images (it takes only 30 s in our computer environment), though the computational cost in the learning phase is relatively high. The decrease in the computational cost remains as a future study.

Although we also implemented an exact EM algorithm which obtains the exact posterior in its E-step, instead of using the hard-EM approximation, we found that the exact EM performed poorly on our segmentation experiments. When applied to the human MRI dataset, the DNN trained by the exact EM predicted most of the voxels as being either background or gray matter. We speculate that this poor performance of the exact EM came from the imbalance in the labels attached to the training datasets. To implement the exact EM after removing such negative effects from the label imbalance remains as another future study.

In this experiment, we assigned pseudo-labels to unlabeled brain images, by simply propagating labels defined on an atlas, directly onto the unlabeled brain images. To assign more reliable pseudo-labels, we can employ sophisticated methods such as mutual registration (Gass, Székely, & Goksel, 2012). The mutual registration considered multiple paths to propagate labels defined on the atlas onto another using the other unlabeled brain images, which is beneficial for reducing the bias to the selected atlas.

We generalized the proposed method such to deal with an arbitrary number of atlases using the ensemble learning approach. We showed the proposed method with 4 atlases outperformed the other methods, including the VoxResNet with 10 atlases, in the evaluation of AVD. However, the DC of the proposed method with any number (less than 10) of atlases was lower than the VoxResNet with 10 atlases. It remains as a future study to improve the generalization of the proposed method to show better performance not only in AVD but also in DC.

Our method thus needed 40% of the annotated images to achieve the comparable performance (in terms of AVD) to what achieved by VoxResNet, in particular for this dataset. This result might not be fully comprehensive due to the inaccessibility to even larger sets of annotated images. However, we also think that these results suggest the possibility to decrease the efforts of manually annotating brain 3D images, which is indeed laborious in our works of large-scale brain image processing.

6. Conclusion

This paper proposed a semi-supervised learning framework to have a DNN-based image registration method, which is trained based not only on a relatively small number of annotated (labeled) images, but also on a relatively large number of unlabeled images. The originally unlabeled images were pseudo-labeled by the label propagation method. Extensive experiments on the human and marmoset brain image datasets showed that our proposed method attained more accurate and stable segmentation than those by the existing registration-based and DNN-based methods. Since the current spatial observation model that represents the mislabeling process is rather simple, there could still be a large room for improving our proposed semi-supervised training algorithm by developing a further sophisticated observation model.

Acknowledgments

This study was supported by the Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS) from Japan Agency for Medical Research and Development (AMED), Post-K Project from Ministry of Education, Sports, Science and Technology (MEXT), Japan, and CREST from Japan Science and Technology Agency (JST).

References

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage*, *26*(3), 839–851.

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.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, *54*(3), 2033–2044.

Bishop, C. M. (2006). *Pattern recognition and machine learning (Information science and statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc..

Borgefors, G. (1986). Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, *34*(3), 344–371.

de Brébisson, A., & Montana, G. (2015). Deep neural networks for anatomical brain segmentation. ArXiv preprint arXiv:1502.02445.

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., & Cuadra, M. B. (2011). A review of atlas-based segmentation for magnetic resonance brain images. *Computer Methods and Programs in Biomedicine*, *104*(3), e158–e177.

Chen, H., Dou, Q., Yu, L., Qin, J., & Heng, P.-A. (2017). VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*.

Gass, T., Székely, G., & Goksel, O. (2012). Semi-supervised segmentation using multiple segmentation hypotheses from a single atlas. In *International MICCAI workshop on medical computer vision* (pp. 29–37). Springer.

Giorgio, A., & De Stefano, N. (2013). Clinical use of brain volumetry. *Journal of Magnetic Resonance Imaging*, *37*(1), 1–14.

Hanbury, A. (2008). A survey of methods for image annotation. *Journal of Visual Languages & Computing*, *19*(5), 617–627.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *European Conference on Computer Vision* (pp. 630–645). Springer.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P., et al. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, *46*(3), 786–802.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Mendrik, A. M., Vincken, K. L., Kuijf, H. J., Breeuwer, M., Bouvy, W. H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A., et al. (2015). MRBrainS Challenge: Online evaluation framework for brain image segmentation in 3T MRI scans. *Computational Intelligence and Neuroscience*, *2015*, 1.

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J., & Išgum, I. (2016). Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Transactions on Medical Imaging*, *35*(5), 1252–1261.

Okano, H., Miyawaki, A., & Kasai, K. (2015). Brain/MINDS: Brain-mapping project in Japan. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *370*(1668), 20140310.

Okano, H., Sasaki, E., Yamamori, T., Iriki, A., Shimogori, T., Yamaguchi, Y., Kasai, K., & Miyawaki, A. (2016). Brain/MINDS: A Japanese national brain project for marmoset neuroscience. *Neuron*, *92*(3), 582–590.

Pai, A., Teng, Y.-C., Blair, J., Kallenberg, M., Dam, E. B., Sommer, S., Igel, C., & Nielsen, M. (2017). Characterization of errors in deep learning-based brain MRI segmentation. In *Deep learning for medical image analysis* (pp. 223–242). Elsevier.

Papandreou, G., Chen, L.-C., Murphy, K. P., & Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1742–1750).

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., & Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, *39*(3), 355–368.

Portela, N. M., Cavalcanti, G. D., & Ren, T. I. (2014). Semi-supervised clustering for MR brain image segmentation. *Expert Systems with Applications*, *41*(4), 1492–1497.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Schindler, K. (2012). An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(11), 4534–4545.

Schnell, S., Saur, D., Kreher, B., Hennig, J., Burkhardt, H., & Kiselev, V. G. (2009). Fully automated classification of HARDI in vivo data using a support vector machine. *NeuroImage*, *46*(3), 642–651.

Smith, R. E., Tournier, J.-D., Calamante, F., & Connelly, A. (2012). Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. *Neuroimage*, *62*(3), 1924–1938.

Tom Vercauteren, Xavier Pennec, A. P., & Ayaches, N. (2009). Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, *45*, S61–S72.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., & Shen, D. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, *108*, 214–224.

Zikic, D., Glocker, B., & Criminisi, A. (2013). Atlas encoding by randomized forests for efficient label propagation. In *International conference on medical image computing and computer-assisted intervention* (pp. 66–73). Springer.