NeuroImage 219 (2020) 117026

Contents lists available at ScienceDirect

NeuroImage

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

AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation

Pierrick Coupé [a,*], Boris Mansencal [a], Michaël Clément [a], Rémi Giraud [b], Baudouin Denis de Senneville [c], Vinh-Thong Ta [a], Vincent Lepetit [a], José V. Manjon [d]

[a] CNRS, Univ. Bordeaux, Bordeaux INP, LABRI, UMR5800, F-33400, Talence, France
[b] Bordeaux INP, Univ. Bordeaux, CNRS, IMS, UMR 5218, F-33400, Talence, France
[c] CNRS, Univ. Bordeaux, IMB, UMR 5251, F-33400, Talence, France
[d] ITACA, Universitat Politècnica de València, 46022, Valencia, Spain

Abstract

Whole brain segmentation of fine-grained structures using deep learning (DL) is a very challenging task since the number of anatomical labels is very high compared to the number of available training images. To address this problem, previous DL methods proposed to use a single convolution neural network (CNN) or few independent CNNs. In this paper, we present a novel ensemble method based on a large number of CNNs processing different overlapping brain areas. Inspired by parliamentary decision-making systems, we propose a framework called AssemblyNet, made of two "assemblies" of U-Nets. Such a parliamentary system is capable of dealing with complex decisions, unseen problem and reaching a relevant consensus. AssemblyNet introduces sharing of knowledge among neighboring U-Nets, an "amendment" procedure made by the second assembly at higher-resolution to refine the decision taken by the first one, and a final decision obtained by majority voting. During our validation, AssemblyNet showed competitive performance compared to state-of-the-art methods such as U-Net, Joint label fusion and SLANT. Moreover, we investigated the scan-rescan consistency and the robustness to disease effects of our method. These experiences demonstrated the reliability of AssemblyNet. Finally, we showed the interest of using semi-supervised learning to improve the performance of our method.

1. Introduction

Quantitative brain analysis is crucial to better understand the human brain and to analyze different brain pathologies. However, whole brain segmentation is still a very challenging problem, mostly due to the high number of anatomical labels compared to the limited number of available training data, especially when considering fine-grained segmentation. Manual segmentation of the whole brain is indeed a very tedious and difficult task, preventing the production of large annotated datasets.

To address this question, several methods have been proposed in the past years. By extending the single-atlas method paradigm, the multi-atlas framework has been successfully applied to whole brain segmentation (Heckemann et al., 2006), (Avants et al., 2011). In such approaches, labeled templates are first nonlinearly registered to the target image. Afterwards, the estimated deformations are applied to the manual segmentations before fusing them. This type of methods efficiently deals with limited training data; however, the required multiple nonlinear registrations can result in a huge computational time. Moreover, regularization involved in registration may prevent to accurately capture complex local anatomical patterns.

To reduce the computational time of multi-atlas methods and to better capture local anatomy, patch-based methods have been introduced (Coupé et al., 2011). In such approaches, the label fusion step is based on the nonlocal patch-based estimator. These methods demonstrated state-of-the-art performance for whole brain segmentation (Wang and Yushkevich, 2013; Rousseau et al., 2011; Asman and Landman, 2013, 2014). One of the main references in the domain is the patch-based joint label fusion (JLF) which won the MICCAI challenge in 2012 (Wang and Yushkevich, 2013) and which is still considered as the state of the art for fine-grained whole brain segmentation. In patch-based methods, usual machine learning such as sparse coding (Tong et al., 2013) or neural networks (Sanroma et al., 2018) has been used in place of the nonlocal estimator. Recently, a fast framework has been proposed (Giraud et al., 2016) to further reduce the computational time required by patch-based methods.

More recently, deep leaning (DL) methods have also been proposed for 3D brain segmentation. Most of these methods were dedicated to coarse segmentation considering only few structures *(e.g.,* <35 structures). For instance QuickNat (Guha Roy et al., 2019) works on 27 structures, Bayesian QuickNat works on 33 structures (Roy et al., 2019), 3DQ works on 28 structures (Paschali et al., 2019), DeepNat works on 25 structures (Wachinger et al., 2018), the method proposed in (Roy et al.,

* Corresponding author.
E-mail address: pierrick.coupe@labri.fr (P. Coupé).

https://doi.org/10.1016/j.neuroimage.2020.117026
Received 20 November 2019; Received in revised form 28 May 2020; Accepted 4 June 2020
Available online 6 June 2020

2018) works on 27 structures and the approach in (Rickmann et al., 2020) works on 32 structures. The problem of whole brain segmentation considering fine-grained structures (*i.e.* >100 structures) is much more complex. Consequently, less works have been dedicated to this problem (de Brebisson and Montana, 2015; Henschel et al., 2019; Huo et al., 2019; Mehta et al., 2017). Moreover, most of the proposed methods were based on 2D frameworks. In fact, due to limited GPU memory, first attempts were based on patch-wise strategies (Wachinger et al., 2018), (de Brebisson and Montana, 2015), (Li et al., 2017) or 2D segmentation (slice by slice) (Guha Roy et al., 2019), (Henschel et al., 2019), (Roy et al., 2017). Only recently, 3D fully convolutional network methods were proposed using reduced input data size (*i.e.*, $128 \times 128 \times 128$ voxels) (Wong et al., 2018) or using an Spatially Localized Atlas Network Tiles (SLANT) strategy (Huo et al., 2019). This latter framework divides the whole volume into overlapping sub-volumes, each one being processed by a different U-Net (Ronneberger et al., 2015) (*e.g.*, 8 or 27). The ensemble SLANT strategy addresses the problem of limited GPU memory and simplifies the complex problem of fine-grained whole brain segmentation into simpler problems, better suited to limited training data.

In this paper, we present a new method able to deal with fine-grained whole brain segmentation at full resolution and based on 3D convolution neural networks (CNNs). To this end, we propose to extend the SLANT framework by using a much larger number of more compact 3D U-Nets (*i.e.*, from 27 to 250) while keeping processing time similar. The main question to address is the optimal organization of this large ensemble of CNNs. To this end, we propose a new framework we call AssemblyNet. Inspired by the decision-making process developed by human societies to deal with complex problems, we decided to model a parliamentary system based on two separate assemblies. Such bicameral – meaning two chambers – parliament has been adopted by many countries around the world. A bicameral system is usually composed of an upper and a lower chamber, both having their own independency to ensure the balance of power. However, an assembly may communicate its vote to the other for amendment. Such parliamentary system is capable of dealing with complex decisions, unseen problem and reaching a relevant consensus. This study extends our conference paper (Coupé et al., 2019) with more complete experiments investigating *i)* the impact of semi-supervised learning *ii)* the scan-rescan reliability of our method and *iii)* the robustness to disease effects of the proposed AssemblyNet. Moreover, we added additional ablation study and a new quality metric. Compared to (Huo et al., 2019), our contributions are: *i)* the use of prior knowledge based on fast atlas registration, *ii)* a knowledge sharing between CNNs using nearest neighbor transfer learning, *iii)* iterative refinement process based on a multiscale cascade of assemblies and *iv)* the use of student-teacher semi-supervised learning based on a well-designed auxiliary dataset.

2. Materials and methods

2.1. Method overview

In AssemblyNet, both assemblies are composed of 3D U-Nets considered as "assembly members" (see Fig. 1). Each member represents one territory (*i.e.*, brain area) in the final vote. To this end, we used spatially localized networks where each U-Net only processes a sub-volume of the global volume, as done in (Huo et al., 2019). Sub-volumes overlap each other, so the final segmentation results from an over complete aggregation of local votes. A majority vote is used to obtain the global segmentation. Moreover, each member can share knowledge with their nearest neighbor in the assembly. In particular, we propose a novel nearest neighbor transfer learning strategy, where weights of the spatially nearest U-Net are used to initialize the next U-Net.

In addition, we also propose to use prior knowledge on the expected final decision which can be viewed as the bill (*i.e.*, draft law) submitted to an assembly for consideration. As prior knowledge, we decided to use nonlinearly registered Atlas prior.

Finally, we also propose modeling communication between both assemblies using an innovative multiscale strategy. In AssemblyNet, we use a multiscale cascade of assemblies where the first assembly produces a coarse decision at $2 \times 2 \times 2$ mm^3. This coarse decision is then transmitted to the second assembly for analysis at $1 \times 1 \times 1$ mm^3. This amendment procedure is similar to an error correction or a refinement step. After consideration by both assemblies, the bill under consideration becomes a law which represents the final segmentation in our system.

2.2. Datasets

Training dataset: 45 T1w MRI from the OASIS dataset (Marcus et al., 2007) manually labeled according to the BrainCOLOR protocol were used for training. The selected images were the same than the ones used in (Huo et al., 2019). The age range was 18-96y for this dataset. This dataset, as provided by Neuromorphometrics, included several pre-processing steps. First, all the MRI scans were corrected for bias field inhomogeneity using (Arnold et al., 2001). Second, all the scans were registered along the anterior commissure (AC) and the posterior commissure (PC) using anatomical landmarks. Therefore, the original space of images was not the native space. The OASIS scan resolution was $1 \times 1 \times 1$ mm^3.

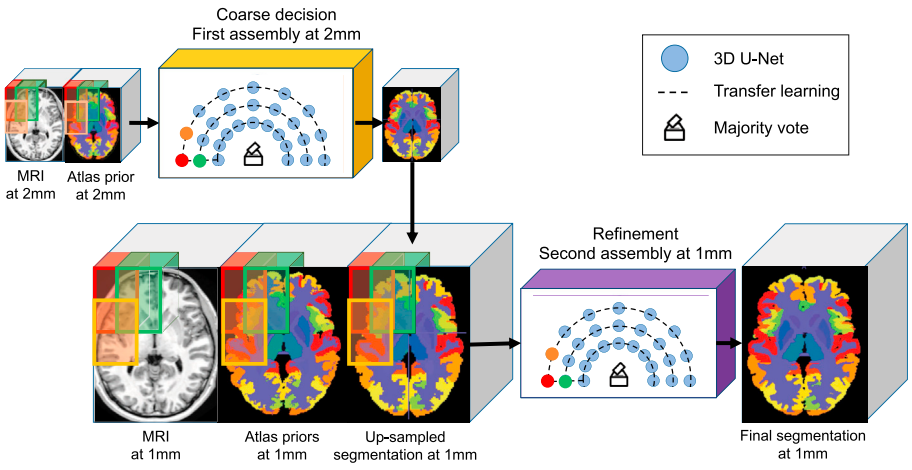

Fig. 1. Illustration of the proposed AssemblyNet framework. Our method is based on two assemblies of 125 3D U-Nets integrated into a multiscale framework. The first assembly (in yellow) provides a coarse segmentation at $2 \times 2 \times 2$ mm^3. The second assembly (in purple) refines this coarse segmentation to produce the final segmentation at $1 \times 1 \times 1$ mm^3. Each 3D U-Net processes a different but overlapping area of the brain. The U-Nets in red in both Assemblies process the area indicated by red rectangles in the input images. The U-Nets in green process the area indicated by green rectangles. During training, the U-Nets in green and orange are initialized using the weights of the U-Nets in red by transfer learning. The output segmentations for each assembly are obtained by majority voting of the 125 3D U-Nets.

Testing dataset: 19 T1w MRI manually labeled according to the BrainCOLOR protocol were used for testing. These MRI came from three different datasets: 5 from the OASIS dataset, the Colin27 atlas (Collins et al., 1998) and 13 from the CANDI database (Kennedy et al., 2012). This testing dataset is the same one used in (Huo et al., 2019). The age range was 20-89y for OASIS, 5-15y for CANDI and the age was 27y for Colin27. The OASIS and CANDI scans as provided by Neuromorphometrics included pre-processing (inhomogeneity correction; AC/PC registration). The resolution of CANDI scans was $0.94 \times 1.5 \times 0.94$ mm^3. The Colin27 atlas was in the MNI space at $0.5 \times 0.5 \times 0.5$ mm^3 (its original space) and included inhomogeneity correction, intensity normalization and averaging of multiple acquisitions.

Scan-Rescan dataset: 8 T1w MRI (from 4 subjects) manually labeled according to the BrainCOLOR protocol were used for scan-rescan experiment. The same expert segmented both the scan and the rescan image. These MRI came from 2 different datasets: 3 from the OASIS dataset (not included in the training) and one from a patient with Alzheimer's disease from the ADNI dataset (Weiner et al., 2013). The OASIS included pre-processing and were in the AC/PC space while ADNI scans were in the native space at $1.2 \times 0.94 \times 0.94$ mm^3.

Pathological dataset: 29 T1w MRI manually labeled according to the BrainCOLOR protocol were included in the pathological dataset. We did not use the rescan image of the AD patient described in the scan-rescan dataset. These MRI came from the ADNI dataset. There were 15 cognitively control subjects and 14 patients with Alzheimer's disease in this dataset. The age of the subjects varied from 62y to 88y. The ADNI scans did not include pre-processing and thus were in their native space.

During our experiments, we used the 132 anatomical labels consistent across subjects (see Tab 1 in supplementary material for the list). In addition to Neuromorphometrics datasets, we used external MRIs from several open access datasets for semi-supervised learning.

Lifespan dataset: 360 unlabeled T1w MRI were randomly selected under constraints from the dataset used in our previous BigData studies (Coupé et al., 2019), (Coupé et al., 2017) to build the lifespan dataset. This dataset was based on 9 datasets publicly available (C-MIND,[1] NDAR,[2] ABIDE,[3] ICBM,[4] IXI,[5] OASIS,[6] AIBL,[7] ADNI1 and ADNI2[8]). From 1y to 90y, we selected 2 females and 2 males for each age (*i.e.*, 2F and 2M of 1 year old, 2F and 2M of 2 years old and so on). Therefore, we obtained a balanced group with 50% of each gender uniformly distributed from 1y to 90y. We made sure that none of the training or testing subjects were selected in this auxiliary dataset.

2.3. AssemblyNet framework

Preprocessing: To homogenize input orientations and intensities, all the images were first preprocessed using the volBrain pipeline[9] (Manjón and Coupé, 2016) with the following steps: *i)* denoising (Manjón et al., 2010), *ii)* inhomogeneity correction (Tustison et al., 2010), *iii)* affine registration into the MNI space ($181 \times 217 \times 181$ voxels at $1 \times 1 \times 1$ mm^3) (Avants et al., 2011), *iv)* tissue-based intensity normalization (Manjón et al., 2008) and *v)* brain extraction (Manjón et al., 2014). Finally, image intensities were centralized and normalized within the brain mask and the background was set to zero.

Atlas priors: To obtain priors knowledge on the expected results, we performed a nonlinear registration of an Atlas (based on the 45 training images) to the subject under consideration. To construct this atlas, we non-linearly registered the training cases to a reference – the MNI152 T1 template – using ANTS software (Avants et al., 2011). The estimated non-linear deformations were then applied to the manual segmentations. Finally, the warped manual segmentations were fused to construct the atlas labels by using a majority vote. The non-linear registration of the Atlas prior to the subject under study was done with an unsupervised deep learning framework similar to VoxelMorph trained on the lifespan dataset (Balakrishnan et al., 2019). There are three main differences between VoxelMorph and our non-linear registration network. First, our network works on deformation fields at 2 mm resolution, which are internally interpolated to 1 mm resolution by the network. This step assumes that the deformation field is spatially smooth. This approach has two 2 benefits, the interpolation process imposes an intrinsic regularization and the smaller size of the volume enables to increase the number of filters at each convolutional layer level giving us almost ten times more learnable parameters (2,165,955 compared to the 259,675 for VoxelMorph). Second, our network is trained using non-skull-striped images contrary to VoxelMorph that requires this pre-processing step. Finally, the used loss function is based on the mean absolute error of both image intensities and labels (last version of VoxelMorph can use also intensities and labels simultaneously, but this last version has been never released as far as we know).

Assembly description: Each assembly was composed of 125 3D U-Nets equally distributed in the MNI space along each axis (*i.e.,* 5 along x, y and z). In the following, we use $U(x, y, z) \; \forall \; x, y, z \in [1..5]$ as the position of the U-Net in the assembly. We experimentally found that $5 \times 5 \times 5$ U-Nets produced the best compromise between segmentation accuracy and computational time (see Results section). Each 3D U-Net processed a sub-volume large enough to ensure at least 50% of overlap between sub-spaces. At the end, a majority vote was used to aggregate the local votes.

Nearest neighbor transfer learning: To enable knowledge sharing between U-Nets within an assembly, we propose a new transfer learning where the weights of a trained U-Net are used to initialize nearest U-Nets in the assembly (*i.e.,* U-Nets processing overlapping sub-volumes and thus dealing with similar anatomy). At the beginning, we trained a first U-Net from scratch in an image corner (*e.g.,* the red U-Net at position *U(1,1,1)* in Fig. 1). The weights of this first trained U-Net were copied to the next U-Net on the same line (*e.g.,* the green U-Net at position *U(2,1,1)* in Fig. 1). This second U-Net was then used to initialize the next network on the same line and so on. Once the first line was trained, each U-Net of the second line *U(x,2,1)* was initialized with the U-Net at the same position on the previous line *U(x,1,1)* and so on (*e.g.,* the orange U-Net was initialized with weights of the red U-Net). Finally, Finally, once all the U-Nets on the same plan *U(x,y,1)* were trained, each U-Net of the next plan *U(x,y,2)* was initialized with the U-Net at the same position on the previous plan *U(x,y,1)* and so on. During the transfer learning, we copied only the weights of the descending/contraction path of the U-Net architecture.

Multiscale cascade of assemblies: To make our decision-making system faster and more robust, we decided to use a multiscale framework. Consequently, the first assembly at $2 \times 2 \times 2$ mm^3 produced a coarse segmentation. Afterwards, an up-sampling to $1 \times 1 \times 1$ mm^3 of this segmentation was performed using nearest neighbor interpolation. The up-sampled segmentation at $1 \times 1 \times 1$ mm^3 of the first assembly was then used as an additional input in the second assembly. Consequently, the second assembly had three 3D sub-volumes as input (*i.e.,* T1w, Atlas priors and up-sampled coarse segmentation all at $1 \times 1 \times 1$ mm^3).

2.4. Ensemble framework

Ensemble is a well-known paradigm in machine learning that is used to improve the global performance of a method. This improvement is obtained by training several models before fusing them. Over

[1] https://research.cchmc.org/c-mind/.
[2] https://ndar.nih.gov.
[3] http://fcon_1000.projects.nitrc.org/indi/abide/.
[4] http://www.loni.usc.edu/ICBM/.
[5] http://brain-development.org/ixi-dataset/.
[6] http://www.oasis-brains.org.
[7] http://adni.loni.ucla.edu/research/protocols/mri-protocols.
[8] www.loni.usc.edu.
[9] https://www.volbrain.upv.es.

the past decades, ensemble learning has been extensively studied, specially to deal with small sample size and complex problems (Wu and Tang, 2019). In brain segmentation, most of the recent DL methods are based on multi-view ensemble (Guha Roy et al., 2019), (Henschel et al., 2019), (Mehta et al., 2017). In such framework, 2D models along axial, coronal and sagittal view are trained before fusing their outputs to enforce 3D consistency. A variant consists in fusing 2D models and 3D models (Mehta et al., 2017), (Zheng et al., 2019). It is also possible to fuse predictions of models working at different scales (Moeskops et al., 2016). Moreover, there exist frameworks based on aggregation of predictions from multiple models with different architectures (Kamnitsas et al., 2018) or trained with different subset of the training dataset (Dolz et al., 2020). In addition, dropout has been proposed to generate several instances of a model at test time to reduce over-fitting (Wu and Tang, 2019). However, all these methods are based on few models (typically <10 networks). The investigation of using a larger number of networks is recent in DL for brain segmentation. In (Huo et al., 2019), SLANT strategy proposes to train more models (>10) on different areas of the brain that enforces model diversity before fusing them. The use of a large ensemble of CNNs has yielded state-of-the-art results for fine-grained whole brain segmentation.

In this paper, the proposed AssemblyNet includes almost all these strategies in a single framework. First, AssemblyNet includes a multiscale cascade of assemblies that allows to take advantage of models trained at different resolutions. Second, each U-Net is trained with a different set for training and validation to take advantage of all the available training data. Third, each U-Net is trained on a different but overlapping brain area that ensures models diversity as in SLANT. In addition, we use dropout at test time to simulate several instances of each U-Net that reduces over-fitting. We also performed temporal averaging of model weights based on snapshot ensembles (Laine and Aila, 2017). Finally, compared to (Huo et al., 2019), we propose to use a much larger number (from 27 to 250) of more compact U-Nets to better deal with limited training data.

2.5. Semi-supervised learning framework

In this study, we also investigated the use of semi-supervised learning (SSL) to further improve segmentation accuracy. SSL aims at using a small amount of labeled data in combination with a large number of unlabeled images to achieve higher performance. In medical image analysis, these techniques are particularly interesting to overcome the limited amount of training data and the complexity of the labeling process.

Over the past years, several strategies have been proposed to make SSL efficient within DL framework mainly based on the teacher-student paradigm. Within this paradigm, SSL methods can integrate a consistency term on the predictions of unlabeled samples to force the student network not to diverge (Laine and Aila, 2017), (Tarvainen et al., 2017). This idea is usually based on a "Mean teacher" by using exponential moving averages over the predictions or the model parameters at different steps during training (Huang et al., 2017). Another teacher-student strategy consists in using the available labeled data to generate weak labels (also called pseudo labels) for unlabeled examples using a teacher model. This weak labels are in turn used as training dataset to improve the robustness of the student network (Yalniz et al., 2019). Afterwards, the student network is fine-tuned on the original labeled data to avoid error propagation.

SSL strategies have been successfully applied to improve segmentation results in different medical applications (Chen et al., 2019; Perone et al., 2019; Zhou et al., 2019). In our context of whole brain segmentation, the authors of (Huo et al., 2019), (Roy et al., 2017) proposed to use auxiliary datasets segmented with traditional tools such as Freesurfer (Fischl, 2012) or non-local spatial staple label fusion (NLSS) (Asman and Landman, 2013) to improve their segmentation framework based on deep learning. However, such methods can require impractical computational burden (e.g., 21 CPU years in (Huo et al., 2019)) and classical tools may provide suboptimal auxiliary segmentations.

Here, we take inspiration of the teacher-student paradigm from (Yalniz et al., 2019) to leverage the fast processing capabilities and high segmentation accuracy of the proposed AssemblyNet. In our SSL framework, an AssemblyNet teacher – trained on the 45 training images – was first used to segment unlabeled images (*i.e.,* the 360 images of the lifespan dataset). Then, these 360 pseudo-labeled images were used to train from scratch an AssemblyNet student. At the end, the AssemblyNet student was fine-tuned on the 45 manual segmentations of the training dataset. As shown in (Yalniz et al., 2019), this fine-tuning step is able to limit error propagation within SSL framework. During our experiments, we investigated the iteration of this procedure considering that the obtained AssemblyNet student could be a good teacher for a second student generation. In our SSL framework, we took care to build the unlabeled images dataset balanced in age and gender in order to limit bias introduction in the pseudo-labeled population.

2.6. Implementation details

Data augmentation: First, the images of the training and lifespan datasets were flipped along mid sagittal plane in the MNI space. Then, we used MixUp data augmentation during training to minimize overfitting problems (Zhang et al., 2017). This method performs a linear interpolation of a random pair of training examples and their corresponding labels.

Training framework: For all the networks, we used the 3D U-Net architecture proposed in (Huo et al., 2019), but with a lower number of filters. Instead of using a basis of 32 filters of $3 \times 3 \times 3$–32 for the first layer, 64 for the second and so on – we selected a basis of 24 filters of $3 \times 3 \times 3$ to reduce by 25% the network size. We experimentally found that this setting reduced memory consumption without impacting performance. The used architecture is presented in Fig. 2. Each block was composed of batch normalization, convolution and ReLU activation. The skip connections between encoder and decoder were based on concatenation. In addition, dropout was done between each level of the encoder. Moreover, upsampling in the decoder was based on trilinear interpolation. Finally, a SoftMax was done before performing argmax to obtain the final label for each voxel.

For all the U-Nets, we used the same parameters: batch size = 1, optimizer = Adam, epoch = 100, loss = Dice and dropout = 0.5 after each block of the descending path. For the U-Nets of the first assembly at $2 \times 2 \times 2\ mm^3$, we used input resolution = $32 \times 48 \times 32$ voxels and input channel = 2 (*i.e.*, T1w and Atlas priors). For the U-Nets of the second assembly at $1 \times 1 \times 1\ mm^3$, we used input resolution = $64 \times 72 \times 64$ voxels and input channel = 3 (*i.e.*, T1w, Atlas priors and up-sampled coarse segmentation). In addition, to compensate for the small batch size, we performed temporal averaging of model weights based on snapshot ensembles (Laine and Aila, 2017). At the end of the 100 epochs, we performed additional 20 epochs where the model estimated at each epoch is averaged with previous ones using a moving average. Such average of model weights along the optimization trajectory leads to better generalization than usual training (Izmailov et al., 2018). For the SSL step, we used only 20 epochs for normal optimization and 10 epochs for moving average. Finally, we also performed dropout at test time (Gal and Ghahramani, 2016). For each U-Net, we generated 3 different outputs before averaging them (with dropout layer active). Such method helps reducing variance of the networks. As in (Huo et al., 2019), the experiments were done with an NVIDIA Titan Xp with 12 GB memory and thus processing times are comparable.

Computational time: The preprocessing steps take around 90 s. The non-linear registration of the atlas takes less than 5 s. The first assembly at $2 \times 2 \times 2\ mm^3$ requires 3 min to segment an image while the second assembly at $1 \times 1 \times 1\ mm^3$ requires 5 min. At the end, the final

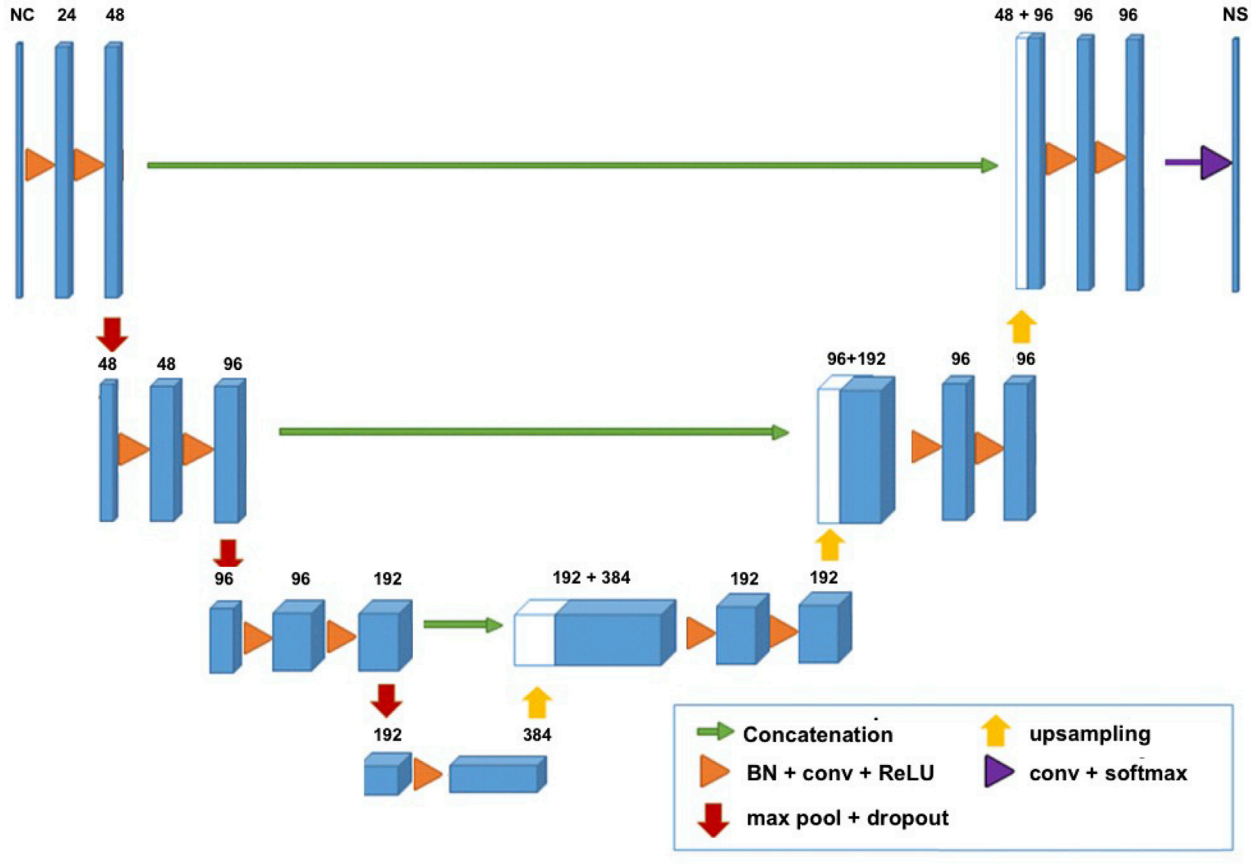

Fig. 2. *Illustration of the used U-Net architecture. The number of input channels (NC) depends on the considered assembly* (i.e., *NC = 2 at 2 × 2 × 2 mm^3 and NC = 3 at 1 × 1 × 1 mm^3). Each block is composed of batch normalization (BN), convolution and ReLU activation. The number of 3 × 3 × 3 filters is indicated on the top of each block. The final output size depends on the number of structures (NS) in the considered sub-volume.*

segmentation is registered back to the original space using the inverse affine transform estimated during preprocessing. This interpolation takes around 30 s. Therefore, the full AssemblyNet process takes around 10 min including preprocessing, segmentation, and inverse registration back to the original space.

2.7. Validation framework

First, for each testing subject, we estimated the average Dice coefficient on the 132 considered anatomical labels (without background) in the original space. Afterwards, we estimated the global mean Dice in % over the 19 images of the testing dataset. In this experiment, we compared AssemblyNet with several state-of-the-art methods. First, the patch-based joint label fusion (JLF) (Wang and Yushkevich, 2013) was used as reference. In addition, we included U-Net (Ronneberger et al., 2015), SLANT-8 and SLANT-27 methods as proposed in (Huo et al., 2019). SLANT-8 is based on 8 U-Nets processing non-overlapping sub--volumes of 86 × 110 × 78 voxels while SLANT-27 is based on 27 U-Nets processing overlapping sub-volumes 96 × 128 × 88 voxels. All these methods were trained on the same 45 training images described in the section datasets. Moreover, all these methods used the following pre-processing steps: *i)* affine registration to MNI space using NiftyReg (Ourselin et al., 2001), *ii)* bias filed correction using N4 (Tustison et al., 2010) and *iii)* intensity harmonization using regression-based normalization. In addition, we included SLANT-27 FT trained on 5111 auxiliary images segmented using NLSS (Asman and Landman, 2014) and fined tuned on the 45 training images. These are the best published results for fine-grained whole brain segmentation to our knowledge. Afterwards, to show the impact of affine registration to MNI space, we also presented the results of "naïve U-Net" working directly in their original space. For all these methods, we report the results published in (Huo et al., 2019). We also used the docker implementation of SLANT-27[10] to produce segmentations of all the considered datasets. Finally, we also compared mean surface distance between AssemblyNet, AssemblyNet SSL (with semi-supervised learning) and SLANT-27 FT.

For the scan-rescan reliability experiment, we rigidly registered the re-scan image into the original space of the scan images (Avants et al., 2011). We then interpolated the re-scan segmentations into the original space of the scan images using the estimated transformation matrix. By this way, we estimated the Dice coefficients between both manual segmentations (*i.e.*, intra-rater consistency) and both automatic segmentations (*i.e.*, intra-method consistency). Moreover, we estimated the method-expert consistency as the Dice coefficients between the automatic segmentation of the rescan images and the manual segmentation of the scan image.

For the experiment on the robustness to disease effects, we first computed the Dice on the 29 ADNI subjects. We then compared the Dice coefficients obtained for cognitively normal (CN) subjects and patients with Alzheimer's Disease (AD) to study the impact of pathology on segmentation accuracy.

For all these experiments, we used one-sided non-parametric Wilcoxon signed-rank test at 95% of confidence to assess the significance of Dice improvement as in (Huo et al., 2019). Moreover, we used one-sided Mann-Whitney rank test at 95% of confidence to assess the significance of Dice decrease between the AD group compared to the CN group.

[10] https://github.com/MASILab/SLANTbrainSeg.

Table 1
Evaluation of the proposed contributions. The mean Dice (std) is evaluated on the 19 images of the test dataset in the original space for the 132 considered labels (without background). Testing time includes image preprocessing and registration back to the original space. * indicates a significant lower Dice compared to AssemblyNet using a Wilcoxon test.

Methods	Atlas prior	Transfer learning	Multi-scale	Dice in % (std)	Training time	Testing time
Assembly at $2 \times 2 \times 2$ mm^3	No	No	–	67.4 (3.4)*	29h	5min
Assembly at $2 \times 2 \times 2$ mm^3	Yes	No	–	67.7 (3.3)*	29h	5min
Assembly at $2 \times 2 \times 2$ mm^3	Yes	Yes	–	67.9 (3.3)*	29h	5min
Assembly at $1 \times 1 \times 1$ mm^3	Yes	Yes	No	72.2 (3.8)*	6 days	7min
AssemblyNet	Yes	Yes	Yes	**73.3 (4.2)**	7 days	10min

Table 2
Impact of the number of U-nets on the Assembly at $2 \times 2 \times 2$ mm^3 with atlas prior and transfer learning. The mean Dice (std) is evaluated on the 19 images of the test dataset in the original space for the 132 considered labels (without background). Testing time includes image preprocessing and registration back to the original space. * indicates a significant lower Dice compared to AssemblyNet based on 343 U-Nets using a Wilcoxon test.

Methods	Number of U-Nets	Dice in % (std)	Training time	Testing time
Assembly at $2 \times 2 \times 2$ mm^3	27 ($3 \times 3 \times 3$)	66.1 (3.4)*	6h	3min
Assembly at $2 \times 2 \times 2$ mm^3	64 ($4 \times 4 \times 4$)	67.6 (3.4)*	15h	3min
Assembly at $2 \times 2 \times 2$ mm^3	125 ($5 \times 5 \times 5$)	**67.9 (3.3)**	29h	5min
Assembly at $2 \times 2 \times 2$ mm^3	216 ($6 \times 6 \times 6$)	**67.9 (3.3)**	2 days	7min
Assembly at $2 \times 2 \times 2$ mm^3	343 ($7 \times 7 \times 7$)	**67.9 (3.3)**	3 days	10min

3. Results

3.1. AssemblyNet performance

First, we evaluated the proposed contributions (see Table 1). Compared to baseline results at $2 \times 2 \times 2$ mm^3 (Dice = 67.4%), the use of Atlas prior provided a gain of 0.3 percentage point (pp) in terms of mean Dice. Moreover, the combination of Atlas prior and transfer learning improved by 0.5 pp the baseline mean Dice. In addition, multiscale cascade of assemblies increased by 1.1 pp the mean Dice obtained with Assembly at $1 \times 1 \times 1$ mm^3 without multiscale cascade (Dice = 72.2%). Finally, AssemblyNet outperformed by 5.9 pp the mean Dice obtained with baseline Assembly at $2 \times 2 \times 2$ mm^3. The Dice coefficients produced by AssemblyNet were significantly better than the Dice coefficients produced by all the considered alternatives. Note these results were obtained using only 45 training cases.

Afterwards, we studied the impact of the number of U-nets on the performance of the Assembly at $2 \times 2 \times 2$ mm^3 with atlas prior and transfer learning (see Table 2). During this experience, the accuracy reached a plateau from 125 U-Nets. Using more networks did not provide additional improvements while increasing computational time. Therefore, in the following, 125 was used as the default number of networks in each assembly.

Table 3
Impact of the proposed semi-supervised learning framework on the 19 images of the testing dataset. The mean Dice (std) is evaluated on the 132 considered labels (without background) in the original space. * indicates a significant lower Dice compared to second generation of AssemblyNet SSL using a Wilcoxon test.

Methods	Training Images	Dice in % (std)	**Training time**	**Library extension time**
AssemblyNet Teacher	45	73.3 (4.2)*	7 days	0s
AssemblyNet SSL First student generation	360	73.6 (4.1)*	12 days	2.5 days
AssemblyNet SSL First student generation	360 + FT 45	73.9 (4.0)	14 days	2.5 days
AssemblyNet SSL Second student generation	360	73.9 (4.0)	26 days	5 days
AssemblyNet SSL Second student generation	360 + FT 45	74.0 (3.9)	28 days	5 days

Table 4
Comparison with state-of-the-art methods on the 19 images of the testing dataset. The mean Dice is evaluated on the 132 considered labels (without background) in the original space. * indicates a significant lower Dice compared to AssemblyNet SSL using a Wilcoxon test when compared to SLANT-27 FT (docker) AssemblyNet and the Assembly at $2 \times 2 \times 2$ mm^3.

Methods	**Training images**	Dice in % (std)	**Training time**	**Testing time**	**Library extension time**
Naïve U-Net (Ronneberger et al., 2015)	45	41.0	33h	1min	0s
U-Net (Huo et al., 2019)	45	57.0	33h	8min	0s
SLANT-8 (Huo et al., 2019)	45	57.0	11 days	10min	0s
JLF (Wang and Yushkevich, 2013)	45	63.4	0s	34h	0s
SLANT-27 (Huo et al., 2019)	45	66.1	42 days	15min	0s
SLANT-27 FT (Huo et al., 2019)	5111 + FT 45	72.9	27 days	15min	21 years[a]
Assembly at $2 \times 2 \times 2$ mm^3	45	67.9 (3.3)*	29h	5min	0s
SLANT-27 FT (docker)	5111 + FT 45	72.6 (2.8)*	27 days	15min	21 years[a]
AssemblyNet	45	73.3 (4.2)*	7 days	10min	0s
AssemblyNet SSL	360 + FT 45	**73.9 (4.0)**	14 days	10min	2.5 days

[a] Library extension time represents the CPU time required to segment 5111 MRI using NLSS (*i.e.*, 34h × 5111). This number of 21 CPU years is reported in (Huo et al., 2019).

3.2. Impact of semi-supervised learning

Second, we evaluated the impact of the proposed teacher-student semi-supervised learning (SSL) framework (see Table 3). The used of the lifespan dataset – labeled with teacher – in the training of the student lead to an improvement of 0.6pp of mean Dice compared to the mean Dice of the teacher (Dice = 73.3%). The fine-tuning (FT) step further increased the mean Dice by 0.3 pp. These results are in line with previous literature (Huo et al., 2019), (Yalniz et al., 2019) on the role of the FT

Table 5
Comparison with state-of-the-art methods using Dice on the different testing datasets (5 adult scans from OASIS, 13 child scans from CANDI child and the high-resolution Colin27 image based on scans average). Using Wilcoxon tests * indicates a significant lower Dice compared to AssemblyNet SSL when compared to SLANT-27 FT (docker) and AssemblyNet.

Methods	Training images	OASIS Dice in % (std)	CANDI Dice in % (std)	Colin27 Dice in %
Naïve U-Net (Ronneberger et al., 2015)	45	60.6 (0.6)	37.5 (4.3)	0.0
U-Net (Huo et al., 2019)	45	70.6 (0.9)	51.4 (8.1)	62.1
SLANT-8 (Huo et al., 2019)	45	69.9 (1.4)	51.9 (7.0)	59.7
JLF (Wang and Yushkevich, 2013)	45	74.6 (0.9)	59.0 (3.3)	64.6
SLANT-27 (Huo et al., 2019)	45	76.6 (0.8)	62.1 (6.2)	66.5
SLANT-27 FT (Huo et al., 2019)	5111 + FT 45	77.6 (1.2)	71.1 (2.3)	73.2
SLANT-27 FT (docker)	5111 + FT 45	75.9 (1.7)*	71.3 (2.2)*	73.5
AssemblyNet	45	78.8 (1.7)	71.1 (2.9)*	74.2
AssemblyNet SSL	360 + FT 45	**79.0** (2.0)	**71.9** (2.9)	**75.0**

step to limit error propagation within semi-supervised learning framework. Finally, the second iteration after FT produced marginal improvement and lead to a mean Dice of 74.0%. This improvement was not significant. Therefore, in the following, we used the first student generation since the time required for the second iteration is not justified by the performance improvement.

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

We compared AssemblyNet with state-of-the-art methods (see Table 4). When considering only methods trained with 45 images, AssemblyNet improved mean Dice obtained with U-Net and SLANT-8 by 16.3 pp, JLF by 9.9 pp and SLANT-27 by 7.2 pp. AssemblyNet was also efficient in terms of training and testing times compared to SLANT-based methods. It has to be noted that the Assembly at $2 \times 2 \times 2$ mm outperformed all the methods using 45 training images (except AssemblyNet) while working at low resolution.

In addition, compared to SLANT-27 FT, AssemblyNet provided better results without library extension while being faster to train and to execute. Using our SSL framework based on 360 + 45 images, AssemblyNet SSL obtained a gain of 1pp compared to SLANT-27 FT trained over 5111 + 45 images. According to (Huo et al., 2019), their library extension required 21 CPU years to be completed. Consequently, such an approach is impractical or very costly using a cloud-based solution. The proposed SSL framework is more practical in term of time and resources. Finally, our AssemblyNet SSL was significantly better than AssemblyNet and SLANT-27 FT (docker). In our framework, SLANT-27 FT (docker) obtained slightly lower results than the ones published in the original paper (Huo et al., 2019). This may come from hardware and environment differences.

In addition, we analyzed the performance of the methods according to the dataset. Mean Dice coefficients obtained on each testing dataset (*i.e.*, OASIS, CANDI and Colin27) are provided in Table 5 (see Figs. 1–4 in supplementary material for boxplots of Dice distributions).

As expected, all the methods performed better on adult scans from the OASIS dataset since the training dataset comes from the same cohort. Moreover, all the images were acquired with the same protocol on the same scanner and provided in the same space. First, we can note the good performance of JLF compared to U-Net and SLANT-8. Moreover, when considering methods trained with 45 training, AssemblyNet outperformed U-Net by 8.2 pp, JLF by 4.4 pp and SLANT-27 by 1.2 pp of mean Dice. When considering all the methods, AssemblyNet SSL obtained significantly better Dice than SLANT-27 FT (docker). It has to be noted that in (Huo et al., 2019), the authors have shown that the SLANT-27 FT significantly outperformed U-Net, JLF and SLANT-8. Finally, on the OASIS images, using SSL did not significantly improve the AssemblyNet results.

Table 6
Comparison with state-of-the-art methods using mean surface distance (MSD) on the different testing datasets (5 adult scans from OASIS, 13 child scans from CANDI child and the high-resolution Colin27 image based on scans average). Using Wilcoxon tests * indicates a significant greater MSD compared to AssemblyNet SSL.

Methods	OASIS MSD in mm	CANDI MSD in mm	Colin27 MSD in mm	Global MSD in mm
SLANT-27 FT (docker)	0.699 (0.076)*	1.062 (0.098)*	0.242	0.923 (0.243)*
AssemblyNet	0.550 (0.058)	1.028 (0.153)*	0.212	0.859 (0.289)*
AssemblyNet SSL	**0.553 (0.067)**	**0.996 (0.134)**	**0.206**	**0.838 (0.270)**

On child scans from the CANDI dataset acquired with a different protocol, we can first note a dramatic drop in performance for all the methods except for AssemblyNet, AssemblyNet SSL and SLANT-27 FT. Moreover, when considering methods trained with 45 training, AssemblyNet outperformed U-Net by 19.7 pp, JLF by 19.2 pp and SLANT-27 by 9 pp of mean Dice. When considering all the methods, AssemblyNet SSL obtained significantly better Dice than AssemblyNet and SLANT-27 FT (docker).

On the high-resolution Colin27 image, we also observed an important decrease of performance for all the methods except for AssemblyNet, AssemblyNet SSL and SLANT-27 FT. As for CANDI dataset, AssemblyNet obtained the best segmentation accuracy with or without SSL on this dataset.

The comparison between Naïve U-Net (working in the original space) and U-Net (working in the MNI space) showed that performing an affine registration to MNI produced a gain of 16 pp (see Table 4). On the OASIS dataset (in the AC/PC space), the training and testing spaces were similar and thus the naïve U-Net obtained descent results (see Table 5). However, on the Colin27 atlas in the MNI space at $0.5 \times 0.5 \times 0.5$ mm^3, the Naïve U-Net obtained a Dice = 0 since this Atlas is not in the training space. While slower, using an additional affine registration to MNI space allowed to improve performance and to be robust to image space and resolution.

Finally, we compared mean surface distance (MSD) from manual segmentations to automatic segmentations on AssemblyNet, AssemblyNet SSL and SLANT-27 FT (see Table 6 and Figs. 5–8 in supplementary material). Average mean surface distance showed similar trends than Dice scores. AssemblyNet SSL produced significant lower MSD in all the considered cases except for OASIS dataset compared to AssemblyNet.

As an illustration, Fig. 3 shows the segmentations of the central slides in the original space obtained by SLANT-27 FT (docker) and AssemblyNet SSL on the first subject of the testing dataset (ID = 1120_3). Both methods provided good segmentations although AssemblyNet SSL segmentation was less smooth especially around sulci (*e.g.*, cerebellum – see red ellipses). Moreover, we can observe an over segmentation of cortical gray matter in SLANT-27 FT segmentation as visible in the error map where structures appeared (see green ellipses). Finally, this figure shows the staircasing artifacts present in the human segmentation (*e.g.*, pallidum – see the pink ellipses) while automatic methods were more regular and consistent.

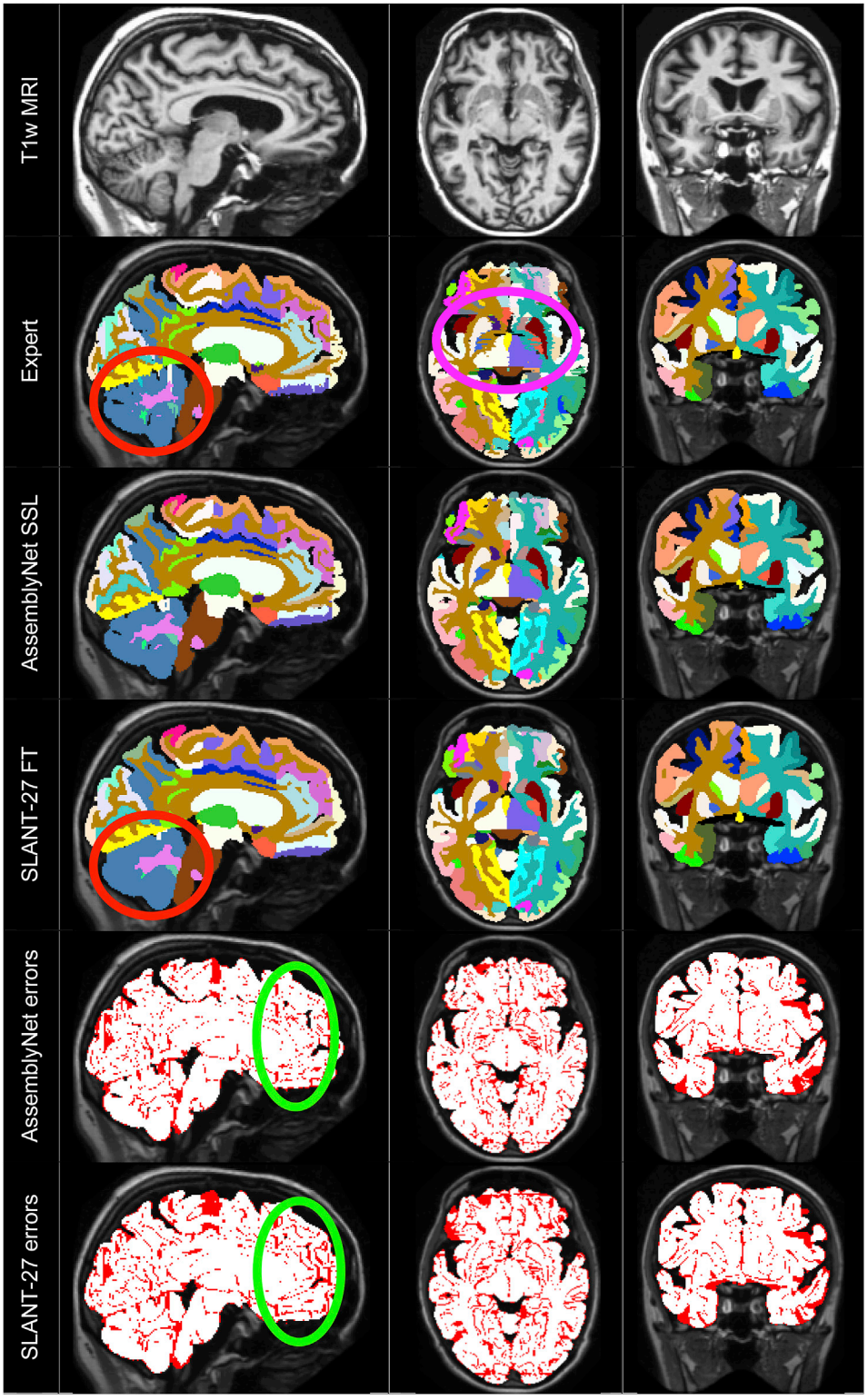

Fig. 3. Example of segmentations in the original space for the first testing subject (ID = 1120_3). First rows: sagittal; axial and coronal views for the T1w MRI. Second row: manual segmentation produced by the expert. Third row: segmentation obtained by our AssemblyNet SSL. Fourth row: segmentation obtained by SLANT-27 FT (docker). Fifth row: binary difference between manual and AssemblyNet SSL segmentations. Last row: binary difference between manual and SLANT-27 FT (docker) segmentations. Colored ellipses indicate areas of interest.

3.4. Scan-rescan consistency

The study of segmentation reproducibility produced by a segmentation method is also highly important especially in medical imaging. Therefore, we carried out a scan-rescan experiment to investigate the consistency and reliability of the proposed method. The results on the 4 images of the scan-rescan dataset are provided in Table 7. It has to be highlighted that there are variations between both acquisitions due to patient's motion, distortion, inhomogeneity and noise. Therefore, agreement between both acquisitions is not expected to be perfect but higher Dice indicates better method stability, consistency and reliability.

First, we estimated the consistency of the segmentations provided by SLANT-27 FT (docker), AssemblyNet, AssemblyNet SSL and the expert on the scan and the rescan images. As expected, we can note that automatic methods were much more consistent (>90% of Dice) than the expert who obtained a scan-rescan segmentation consistency of 76.8% of Dice.

Table 7
Reliability study on the scan-rescan datasets (3 adult scans from OASIS and one scan of a patient with AD from ADNI). The intra-method consistency is the mean Dice between the automatic segmentations obtained on the scan and rescan images. The Expert-Method consistency is the mean Dice between the automatic segmentation on the rescan image and the manual segmentation of the scan image. The Dice coefficients obtained on the scan and the rescan images are averaged. The intra-rater consistency is the mean Dice between the manual segmentations obtained on the scan and rescan images. Using Wilcoxon tests * indicates a significant lower Dice compared to AssemblyNet SSL.

Methods	Training images	Intra-method consistency Dice in % (std)	Intra-rater consistency Dice in % (std)	Expert-Method consistency Dice in % (std)
SLANT-27 FT (docker)	5111 + FT 45	91.7 (1.5)*	76.8 (4.1)	72.9 (1.7)*
AssemblyNet	45	92.0 (1.8)*		75.6 (1.7)
AssemblyNet SSL	360 + FT 45	**92.8 (1.8)**		**75.8 (1.9)**

Moreover, we can see that AssemblyNet was more consistent than SLANT-27 FT especially when using the proposed teacher-student SSL framework. The method consistency of AssemblyNet SSL was significantly higher than the SLANT-27 FT (docker) and AssemblyNet.

In addition, we estimated the Expert-Method consistency as the mean Dice coefficients between the automatic segmentation on the rescan image and the manual segmentation of the scan image. The Expert-Method consistency of AssemblyNet SSL was significantly higher than the consistency of SLANT-27 FT (docker) but not than the consistency of AssemblyNet. Finally, the Expert-Method consistencies obtained by automatic methods were not significantly lower than the intra-expert consistency although this difference was almost significant for SLANT-27 FT docker ($p = 0.36$ for AssemblyNet SSL, $p = 0.23$ for AssemblyNet and $p = 0.07$ for SLANT-27 FT).

3.5. Robustness to disease effects

The last part of our validation is dedicated to robustness to disease effects. To this end, we compared SLANT-27 FT (docker), AssemblyNet and AssemblyNet SSL on the pathological dataset composed of CN subjects and AD patients.

First, we estimated the mean Dice for both groups (see Table 8). For the CN group, AssemblyNet SSL obtained a significantly better Dice than both other methods. For the AD group, we obtained similar results although the improvement obtained by AssemblyNet SSL was higher for AD group (2.2 pp) than for the CN group (1.7 pp) compared to SLANT-27 FT (docker).

In addition, we compared the accuracy between CN and AD groups for the three methods. We found no significant differences between groups for all the methods although this difference was almost significant for SLANT-27 FT docker ($p = 0.31$ for AssemblyNet SSL, $p = 0.18$ for AssemblyNet and $p = 0.06$ for SLANT-27 FT).

Finally, AssemblyNet obtained a global Dice 73.1% without SSL and 73.6% with SSL, while SLANT-27 FT obtained 71.6%. The results of AssemblyNet SSL were significantly better than the results obtained with both other methods.

4. Discussion

In this work, we presented a novel whole brain segmentation framework based on a large number of 3D CNN (*i.e.*, 250 U-Nets) called AssemblyNet. First, we showed that the use of Atlas prior, nearest neighbor transfer learning and multiscale cascade of Assemblies enable to improve global segmentation accuracy. In further work alternative

Table 8
Methods comparison on the pathological dataset (29 scans from the ADNI dataset including 15 CN and 14 patients with AD). * indicates a significant lower Dice compared to AssemblyNet SSL using Wilcoxon tests.

Methods	Training images	CN Dice in % (std)	AD Dice in % (std)	ADNI Dice in % (std)
SLANT-27 FT (docker)	5111 + FT 45	72.3 (1.6)*	71.0 (2.6)*	71.6 (2.2)*
AssemblyNet	45	73.6 (1.6)*	72.6 (2.6)*	73.1 (2.2)*
AssemblyNet SSL	360 + FT 45	**74.0 (1.5)**	**73.2 (2.5)**	**73.6 (2.1)**

options could be investigated. First, atlas prior could be replaced by fast multi-atlas prior. Thanks to nonlinear registration methods based on deep learning, such prior is no more too expensive. Moreover, more advanced communication between assembly members should be investigated. Recent advances in multi-agent reinforcement learning seem a promising way (Omidshafiei et al., 2017). Finally, in this paper, we focused only on the optimal organization of a large group of CNNs without studying the optimal assembly composition. Additional works should investigate this point, for instance by introducing model diversity in the assembly.

Second, we studied the impact of the proposed SSL based on a teacher-student paradigm. We showed that using few hundreds of well-balanced unlabeled data could significantly improve the results of AssemblyNet in all the cases (*i.e.,* unseen acquisition protocol, age period and pathology). Compared to previous methods using larger auxiliary datasets labeled with classical tools (Huo et al., 2019), (Roy et al., 2017), the proposed SSL framework is more practical in terms of computational time and resources. However, SSL is currently receiving special attention in deep learning community. Consequently new paradigms should be considered (Ren et al., 2018), (Luo et al., 2018).

Afterwards, we compared our AssemblyNet with state-of-the-art methods. We demonstrated the high performance of our method in terms of segmentation accuracy and computational time. First, these experiments demonstrated the advantage of using several CNNs to segment the whole brain since SLANT-27 and AssemblyNet clearly outperformed the use of a single U-Net. Moreover, these results showed that using a larger number of simpler CNNs within a multiscale framework is an efficient strategy. While for OASIS we obtained Dice higher than intra-expert consistency, the accuracy on the CANDI dataset is still limited. This can come from several factors such as the lower image quality (*e.g.,* more motion artifacts in child images) or the larger distance between adult training dataset and this child dataset. These points should be deeper investigated in future whole brain segmentation methods. In terms of computational time, our method could be further improved by using several GPUs. At training time, once the first U-Net is trained, several following U-Nets can be trained in parallel despite our transfer learning strategy. At testing time, AssemblyNet can be fully parallelized and thus the processing time could be drastically reduced using multiple GPUs.

In addition, we investigated the scan-rescan consistency of the proposed method. We showed that the intra-method consistency of our method reached 92.8% while intra-rater consistency was limited to 76.8%. This result clearly demonstrates that our automatic method segments the whole brain in a more consistent manner than a human expert. Moreover, we show that the expert-method consistency obtained with our method is not significantly lower than the intra-rater consistency, which is an encouraging result. However, these results were obtained using only 4 scan-rescan subjects. In addition, the Neuromorphometrics dataset does not contain material to evaluate method reproducibility (same image segmented twice by the same expert). Such data would have been useful to evaluate if automatic methods had reached human

variability. Finally, these results raise the question of using human segmentation as "gold standard" with a consistency lower than 80%. Semi-manual segmentations could be considered in order to reduce intra-rater variability.

Finally, we studied the robustness of AssemblyNet to pathology. To this end, we compared its accuracy on CN and AD groups. We observed a small but non-significant decrease of Dice for the AD group compared to CN group. Moreover, compared to SLANT-27, AssemblyNet was less impacted by the presence of the pathology. This is a first step towards a more extensive validation with other pathologies.

5. Conclusion

In this paper, we proposed to use a large number of CNNs to perform whole brain segmentation. We investigated how to organize this large ensemble of CNNs to accurately segment the brain. To this end, we designed a novel deep decision-making process called AssemblyNet based on two assemblies of 125 3D U-Nets. Our validation showed the very competitive results of AssemblyNet compared to state-of-the-art methods. We also demonstrated that AssemblyNet is very efficient to deal with limited training data and to accurately achieve segmentation in a practical training and testing times. Finally, we demonstrated the interest of semi-supervised learning to improve the performance of our method on unseen acquisition protocol, age period and pathology.

CRediT authorship contribution statement

Pierrick Coupé: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition. **Boris Mansencal:** Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Michaël Clément:** Writing - original draft, Writing - review & editing. **Rémi Giraud:** Writing - original draft, Writing - review & editing. **Baudouin Denis de Senneville:** Writing - original draft, Writing - review & editing. **Vinh-Thong Ta:** Writing - original draft, Writing - review & editing. **Vincent Lepetit:** Writing - original draft, Writing - review & editing. **José V. Manjon:** Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition.

Acknowledgements

This work benefited from the support of the project DeepvolBrain of the French National Research Agency (ANR-18-CE45-0013). This study was achieved within the context of the Laboratory of Excellence TRAIL ANR-10-LABX-57 for the BigDataBrain project. Moreover, we thank the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX- 03-02, HL-MRI Project), Cluster of excellence CPU and the CNRS/INSERM for the DeepMultiBrain project. This study has been also supported by the DPI2017-87743-R grant from the Spanish Ministerio de Economia, Industria Competitividad. The authors gratefully acknowledge the support of NVIDIA Corporation with their donation of the TITAN Xp GPU used in this research.

Moreover, this work is based on multiple samples. We wish to thank all investigators of these projects who collected these datasets and made them freely accessible.

The C-MIND data used in the preparation of this article were obtained from the C-MIND Data Repository (accessed in Feb 2015) created by the C-MIND study of Normal Brain Development. This is a multisite, longitudinal study of typically developing children from ages newborn through young adulthood conducted by Cincinnati Children's Hospital Medical Center and UCLA and supported by the National Institute of Child Health and Human Development (Contract #s HHSN275200900018C). A listing of the participating sites and a complete listing of the study investigators can be found at https://research.cchmc.org/c-mind.

The NDAR data used in the preparation of this manuscript were obtained from the NIH-supported National Database for Autism Research (NDAR). NDAR is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in autism. The NDAR dataset includes data from the NIH Pediatric MRI Data Repository created by the NIH MRI Study of Normal Brain Development. This is a multisite, longitudinal study of typically developing children from ages newborn through young adulthood conducted by the Brain Development Cooperative Group and supported by the National Institute of Child Health and Human Development, the National Institute on Drug Abuse, the National Institute of Mental Health, and the National Institute of Neurological Disorders and Stroke (Contract #s N01- HD02-3343, N01-MH9-0002, and N01-NS-9-2314, -2315, -2316, -2317, -2319 and -2320). A listing of the participating sites and a complete listing of the study investigators can be found at http://pediatricmri.nih.gov/nihpd/info/participating_centers.html.

The ADNI data used in the preparation of this manuscript were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). The ADNI is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics NV, Johnson & Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffmann-La Roche, Schering-Plough, Synarc Inc., as well as nonprofit partners, the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to the ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of California, Los Angeles. This research was also supported by NIH grants P30AG010129, K01 AG030514 and the Dana Foundation.

The OASIS data used in the preparation of this manuscript were obtained from the OASIS project funded by grants P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584. See http://www.oasis-brains.org/ for more details.

The AIBL data used in the preparation of this manuscript were obtained from the AIBL study of ageing funded by the Common-wealth Scientific Industrial Research Organization (CSIRO; a publicly funded government research organization), Science Industry Endowment Fund, National Health and Medical Research Council of Australia (project grant 1011689), Alzheimer's Association, Alzheimer's Drug Discovery Foundation, and an anonymous foundation. See www.aibl.csiro.au for further details.

The ICBM data used in the preparation of this manuscript were supported by Human Brain Project grant PO1MHO52176-11 (ICBM, P.I. Dr John Mazziotta) and Canadian Institutes of Health Research grant MOP-34996.

The IXI data used in the preparation of this manuscript were supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) GR/S21533/02 - http://www.brain-development.org/.

The ABIDE data used in the preparation of this manuscript were supported by ABIDE funding resources listed at http://fcon_1000.projects.nitrc.org/indi/abide/. ABIDE primary support for the work by Adriana Di Martino was provided by the NIMH (K23MH087770) and the Leon Levy Foundation. Primary support for the work by Michael P. Milham and the INDI team was provided by gifts from Joseph P. Healy and the Stavros Niarchos Foundation to the Child Mind Institute, as well as by an NIMH award to MPM (R03MH096321). http://fcon_1000.projects.nitrc.org/indi/abide/

This manuscript reflects the views of the authors and may not reflect the opinions or views of the database providers.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.117026.

References

Arnold, J.B., et al., 2001. Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. May Neuroimage 13 (5), 931–943. https://doi.org/10.1006/nimg.2001.0756.

Asman, A.J., Landman, B.A., 2013. Non-local statistical label fusion for multi-atlas segmentation. Feb Med. Image Anal. 17 (2), 194–208. https://doi.org/10.1016/j.media.2012.10.002.

Asman, A.J., Landman, B.A., 2014. Hierarchical performance estimation in the statistical label fusion framework. Oct Med. Image Anal. 18 (7), 1070–1081. https://doi.org/10.1016/j.media.2014.06.005.

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. Feb Neuroimage 54 (3), 2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE TMI*. https://doi.org/10.1109/TMI.2019.2897538, 1–1.

Chen, S., Bortsova, G., García-Uceda Juárez, A., van Tulder, G., de Bruijne, M., 2019. Multi-task attention-based semi-supervised learning for medical image segmentation. In: *Medical Image Computing And Computer Assisted Intervention – MICCAI 2019*, Cham, pp. 457–465. https://doi.org/10.1007/978-3-030-32248-9_51.

Collins, D.L., et al., 1998. Design and construction of a realistic digital brain phantom. IEEE TMI 17 (3), 463–468. https://doi.org/10.1109/42.712135.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Jan Neuroimage 54 (2), 940–954. https://doi.org/10.1016/j.neuroimage.2010.09.018.

Coupé, P., Catheline, G., Lanuza, E., Manjón, J.V., 2017. Towards a unified analysis of brain maturation and aging across the entire lifespan: a MRI analysis. Hum. Brain Mapp. 38 (11), 5501–5518. https://doi.org/10.1002/hbm.23743.

Coupé, P., Manjón, J.V., Lanuza, E., Catheline, G., 2019. "Lifespan changes of the human brain in Alzheimer's disease. Mar Sci. Rep. 9 (1), 3998. https://doi.org/10.1038/s41598-019-39809-8.

Coupé, P., et al., 2019. AssemblyNet: a novel deep decision-making process for whole brain MRI segmentation. MICCAI. https://doi.org/10.1007/978-3-030-32248-9_52.

de Brebisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: IEEE CVPR Workshops, pp. 20–28.

Dolz, J., Desrosiers, C., Wang, L., Yuan, J., Shen, D., Ben Ayed, I., 2020. Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. Jan Comput. Med. Imag. Graph. 79, 101660. https://doi.org/10.1016/j.compmedimag.2019.101660.

Fischl, B., 2012. FreeSurfer. Aug Neuro*image* 62 (2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021.

Gal, Y., Ghahramani, Z., 2016. A theoretically grounded application of dropout in recurrent neural networks. Adv. Neural Inf. Process. Syst. 1019–1027.

Giraud, R., Ta, V.-T., Papadakis, N., Manjón, J.V., Collins, D.L., Coupé, P., 2016. An Optimized PatchMatch for multi-scale and multi-feature label fusion. Jan Neuroimage 124, 770–782. https://doi.org/10.1016/j.neuroimage.2015.07.076.

Guha Roy, A., Conjeti, S., Navab, N., Wachinger, C., QuickNAT, ", 2019. A fully convolutional network for quick and accurate segmentation of neuroanatomy. Feb Neuroimage 186, 713–727. https://doi.org/10.1016/j.neuroimage.2018.11.042.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Oct Neuroimage 33 (1), 115–126. https://doi.org/10.1016/j.neuroimage.2006.05.061.

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M., 2019. FastSurfer–A Fast and Accurate Deep Learning Based Neuroimaging Pipeline. 1910.03866.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q., 2017. Snapshot Ensembles: Train 1, Get M for Free. Mar [Online]. Available: 1704.00109 [cs]. (Accessed 28 April 2020) http://arxiv.org/abs/1704.00109.

Huo, Y., et al., 2019. 3D whole brain segmentation using spatially localized atlas network tiles. Jul Neuroimage 194, 105–119. https://doi.org/10.1016/j.neuroimage.2019.03.041.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G., 2018. Averaging Weights Leads to Wider Optima and Better Generalization. Mar [Online]. Available: 1803.05407. (Accessed 15 March 2019) http://arxiv.org/abs/1803.05407.

Kamnitsas, K., et al., 2018. Ensembles of multiple models and architectures for robust brain tumour segmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Cham, pp. 450–462. https://doi.org/10.1007/978-3-319-75238-9_38.

Kennedy, D.N., Haselgrove, C., Hodge, S.M., Rane, P.S., Makris, N., Frazier, J.A., 2012. CANDIShare: a resource for pediatric neuroimaging data. Neuroinformatics 10 (3), 319–322. https://doi.org/10.1007/s12021-011-9133-y.

Laine, S., Aila, T., 2017. Temporal Ensembling for Semi-supervised Learning. Mar [Online]. Available: 1610.02242 [cs]. (Accessed 13 November 2019) http://arxiv.org/abs/1610.02242.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T., 2017. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. In: Information Processing in Medical Imaging, Cham, pp. 348–360. https://doi.org/10.1007/978-3-319-59050-9_28.

Luo, Y., Zhu, J., Li, M., Ren, Y., Zhang, B., Mar. 2018. Smooth neighbors on teacher graphs for semi-supervised learning [Online]. Available: *1711.00258 [cs, stat]*. (Accessed 13 November 2019) http://arxiv.org/abs/1711.00258.

Manjón, J.V., Coupé, P., 2016. volBrain: an online MRI brain volumetry system. Front. Neuroinf. 10, 30.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M., 2010. "Adaptive non-local means denoising of MR images with spatially varying noise levels. JMRI 31 (1), 192–203.

Manjón, J.V., Eskildsen, S.F., Coupé, P., Romero, J.E., Collins, D.L., Robles, M., 2014. Nonlocal intracranial cavity extraction. IJBI 2014, 10.

Manjón, J.V., et al., 2008. Robust MRI brain tissue parameter estimation by multistage outlier rejection. Magn. Reson. Med.: An Official Journal of the International Society for Magnetic Resonance in Medicine 59 (4), 866–873.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cognit. Neurosci. 19 (9), 1498–1507.

Mehta, R., Majumdar, A., Sivaswamy, J., 2017. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. J. Med. Imag. 4 (2), 024003.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J.N.L., Išgum, I., 2016. Automatic segmentation of MR brain images with a convolutional neural network. May IEEE Trans. Med. Imag. 35 (5), 1252–1261. https://doi.org/10.1109/TMI.2016.2548501.

Omidshafiei, S., Pazis, J., Amato, C., How, J.P., Vian, J., 2017. Deep Decentralized Multi-Task Multi-Agent Reinforcement Learning under Partial Observability. Jul [Online]. Available: 1703.06182 [cs]. (Accessed 13 November 2019) http://arxiv.org/abs/1703.06182.

Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N., Jan. 2001. Reconstructing a 3D structure from serial histological sections. Image Vis Comput. 19 (1), 25–31. https://doi.org/10.1016/S0262-8856(00)00052-4.

Paschali, M., Gasperini, S., Roy, A.G., Fang, M.Y.-S., Navab, N., 2019. 3DQ: compact quantized neural networks for volumetric whole brain segmentation. In: *Medical Image Computing And Computer Assisted Intervention – MICCAI 2019*, Cham, pp. 438–446. https://doi.org/10.1007/978-3-030-32248-9_49.

Perone, C.S., Ballester, P., Barros, R.C., Cohen-Adad, J., 2019. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. Jul Neuroimage 194, 1–11. https://doi.org/10.1016/j.neuroimage.2019.03.026.

Ren, M., et al., 2018. Meta-Learning for Semi-supervised Few-Shot Classification. Mar [Online]. Available: 1803.00676 [cs, stat]. (Accessed 13 November 2019) http://arxiv.org/abs/1803.00676.

Rickmann, A.-M., Roy, A.G., Sarasua, I., Wachinger, C., 2020. Recalibrating 3D ConvNets with project excite. IEEE Trans. Med. Imag. https://doi.org/10.1109/TMI.2020.2972059, 1–1.

Ronneberger, O., Fischer, P., Brox, T., U-Net, ", 2015. Convolutional networks for biomedical image segmentation. In: MICCAI 2015, pp. 234–241.

Rousseau, F., Habas, P.A., Studholme, C., 2011. A supervised patch-based approach for human brain labeling. Oct IEEE Trans. Med. Imag. 30 (10), 1852–1862. https://doi.org/10.1109/TMI.2011.2156806.

Roy, A.G., Conjeti, S., Sheet, D., Katouzian, A., Navab, N., Wachinger, C., 2017. Error corrective boosting for learning fully convolutional networks with limited data. In: MICCAI 2017, pp. 231–239.

Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: *Medical Image Computing And Computer Assisted Intervention – MICCAI 2018*, Cham, pp. 421–429. https://doi.org/10.1007/978-3-030-00928-1_48.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., 2019. Bayesian QuickNAT: model uncertainty in deep whole-brain segmentation for structure-wise quality control. Jul Neuroimage 195, 11–22. https://doi.org/10.1016/j.neuroimage.2019.03.042.

Sanroma, G., et al., 2018. Learning non-linear patch embeddings with neural networks for label fusion. Feb Med. Image Anal. 44, 143–155. https://doi.org/10.1016/j.media.2017.11.013.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances In Neural Information Processing Systems, vol. *30*. Curran Associates, Inc., pp. 1195–1204

Tong, T., Wolz, R., Coupé, P., Hajnal, J.V., Rueckert, D., 2013. Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. Aug Neuroimage 76, 11–23. https://doi.org/10.1016/j.neuroimage.2013.02.069.

Tustison, N.J., et al., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imag. 29 (6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.

Wachinger, C., Reuter, M., Klein, T., DeepNAT, ", 2018. Deep convolutional neural network for segmenting neuroanatomy. Apr Neuroimage 170, 434–445. https://doi.org/10.1016/j.neuroimage.2017.02.035.

Wang, H., Yushkevich, P., 2013. "Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation. Front. Neuroinf. 7 https://doi.org/10.3389/fninf.2013.00027.

Weiner, M.W., et al., 2013. "The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Sep Alzheimer's Dementia 9 (5), e111–e194. https://doi.org/10.1016/j.jalz.2013.05.1769.

Wong, K.C.L., Moradi, M., Tang, H., Syeda-Mahmood, T., 2018. 3D segmentation with exponential logarithmic loss for highly unbalanced object sizes. In: MICCAI 2018, pp. 612–619.

Wu, J., Tang, X., 2019. Brain Segmentation Based on Multi-Atlas Guided 3D Fully Convolutional Network Ensembles. Jan [Online]. Available: 1901.01381 [cs]. (Accessed 23 April 2020) http://arxiv.org/abs/1901.01381.

Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D., 2019. Billion-scale Semi-supervised Learning for Image Classification. May [Online]. Available: 1905.00546 [cs]. (Accessed 29 October 2019) http://arxiv.org/abs/1905.00546.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: beyond Empirical Risk Minimization [Online]. Available: 1710.09412. (Accessed 14 March 2019) http://arxiv.org/abs/1710.09412.

Zheng, H., et al., 2019. A new ensemble learning framework for 3D biomedical image segmentation. presented at the Proceedings of the AAAI Conference on Artificial Intelligence 33, 5909–5916.

Zhou, Y., et al., 2019. Collaborative learning of semi-supervised segmentation and classification for medical images,. Jun. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2074–2083. https://doi.org/10.1109/CVPR.2019.00218.