3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study

Jose Dolz[a,*], Christian Desrosiers[a], Ismail Ben Ayed[a]

[a] *LIVIA Laboratory, École de technologie supérieure (ETS), Montreal, QC, Canada*

Abstract

This study investigates a 3D and fully convolutional neural network (CNN) for subcortical brain structure segmentation in MRI. 3D CNN architectures have been generally avoided due to their computational and memory requirements during inference. We address the problem via small kernels, allowing deeper architectures. We further model both local and global context by embedding intermediate-layer outputs in the final prediction, which encourages consistency between features extracted at different scales and embeds fine-grained information directly in the segmentation process. Our model is efficiently trained end-to-end on a graphics processing unit (GPU), in a single stage, exploiting the dense inference capabilities of fully CNNs.

We performed comprehensive experiments over two publicly available data sets. First, we demonstrate a state-of-the-art performance on the ISBR dataset. Then, we report a *large-scale* multi-site evaluation over 1112 unregistered subject data sets acquired from 17 different sites (ABIDE data set), with ages ranging from 7 to 64 years, showing that our method is robust to various acquisition protocols, demographics and clinical factors. Our method yielded segmentations that are highly consistent with a standard atlas-based approach, while running in a fraction of the time needed by atlas-based methods and avoiding registration/normalization steps. This makes it convenient for massive multi-site neuroanatomical imaging studies. To the best of our knowledge, our work is the first to study subcortical structure segmentation on such large-scale and heterogeneous data.

Keywords: Deep learning, MRI segmentation, brain, 3D CNN, fully CNN.

1. Introduction

Accurate segmentation of subcortical brain structures is crucial to the study of a breadth of brain disorders, e.g., schizophrenia [1], Parkinson [2], autism [3], multiple-sclerosis lesions [4, 5], and to the assessment of structural brain abnormalities [6]. For instance, changes in the morphology and developmental

[*] Corresponding author: jose.dolz.upv@gmail.com

trajectories of the caudate nucleus, putamen and nucleus accumbens have been associated with autism spectrum disorder (ASD), and may be linked to the occurrence of restricted and repetitive behaviors [7]. Accurate segmentation of these structures would help understanding such complex disorders, monitoring their progression and evaluating treatment outcomes.

Automating subcortical structure segmentation remains challenging, despite the substantial research interest and efforts devoted to this computational problem. Clinicians still rely on manual delineations, a prohibitively time-consuming process, which depends on rater variability, leading to substantial inconsistencies in the segmentations [8]. These issues impede the use of manual segmentation for very large data sets, such as those currently used in various multi-center neuroimaging studies. Therefore, there is a critical need for fast, accurate, reproducible, and fully automated methods for segmenting subcortical brain structures.

1.1. Prior art

A multitude of (semi-) automatic methods have been proposed for segmenting brain structures [9]. We can divide prior-art methods into four main categories: atlas-based methods [10, 11], statistical models [12, 13], deformable models [14] and machine learning based classifiers [15, 16]. Although atlas-based methods often provide satisfactory results, segmentation times are typically long (ranging from several minutes to hours) due to complex registration steps. Furthermore, atlas-based methods may not be able to capture the full anatomical variability of target subjects (e.g., subjects of young age or with structural abnormalities), and can fail in cases of large misalignments or deformations. Unlike atlas-based methods, statistical approaches use training data to learn a parametric model describing the variability of specific brain structures (e.g., shapes, textures, etc.). When the number of training images is small compared to the number of parameters to learn, these approaches might result in overfitting the data, thereby introducing biases in the results. The robustness of such statistical approaches might also be affected by the presence of noise in training data. Finally, because parameters are updated iteratively by searching in the vicinity of the current solution, an accurate initialization is required for such approaches to converge to the correct structure. Deformable models do not need training data, nor prior knowledge, unlike statistical models. Because they can evolve to fit any target structure, such models are considered to be highly flexible compared to other segmentation methods. Yet, deformable models are quite sensitive to the initialization of the deformed contour and the deformation stopping criteria, both of which depend on the characteristics of the problem. The last category of methods, based on machine learning, uses training images to learn a predictive model that assigns class probabilities to each pixel/voxel. These probabilities are sometimes used as unary potentials in standard regularization techniques such as graph cuts [17]. Recently, machine learning approaches have achieved state-of-the-art performances in segmenting brain structures [9, 15]. Nevertheless, these approaches usually involve heavy algorithm design, with carefully engineered, application-dependent features and meta-parameters, which limit their applicability to different brain structures and modalities.

Deep learning has recently emerged as a powerful tool, achieving state-of-the art results in numerous applications of pattern or speech recognition. Unlike traditional methods that use hand-crafted features, deep learning techniques have the ability to learn hierarchical features representing different levels of abstraction, in a data-driven manner. Among the different types of deep learning approaches, convolutional neural networks (CNNs) [18, 19] have shown the greatest potential for computer vision and image analysis problems. In biomedical imaging, CNNs have been recently investigated for several neuroimaging applications [20, 21, 22, 23]. For instance, Ciresan et al. [20] used a CNN to accurately segment neuronal membranes in electron microscopy images. In this study, a sliding-window strategy was applied to predict the class probabilities of each pixel, using patches centered at the pixels as input to the network. An important drawback of this strategy is that its label prediction is based on very localized information. Moreover, since the prediction must be carried out for each pixel, this strategy is typically slow. Zhang et al. [21] presented a CNN method to segment three brain tissues (white matter, gray matter and cerebrospinal fluid) from multi-sequence magnetic resonance imaging (MRI) images of infants (6- to 8-month old). As inputs to the network, 2D images corresponding to a single plane were used. Deep CNNs were also investigated for glioblastoma tumor segmentation [22], using an architecture with several pathways, which modeled both local and global-context features. Pereira et al. [23] presented a different CNN architecture for segmenting brain tumors in MRI data, exploring the use of small convolution kernels.

Several recent studies investigated CNNs for segmenting subcortical brain structures [17, 24, 25, 26, 27]. For instance, Lee et al. [24] presented a CNN-based approach to learn discriminative features from expert-labelled MR images. The study in [25] used CNNs to segment brain structures in images from five different datasets, and reported a performance evaluation for subjects in various age groups, ranging from pre-term infants to older adults. A multi-scale patch-based strategy was used to improve the results, where patches of different sizes were extracted around each pixel as input to the network. Although medical images are often in the form of 3D volumes (e.g., MRI or computed tomography scans), most of the existing CNN approaches used a slice-by-slice analysis of 2D images. An obvious advantage of a 2D approach, compared to one using 3D images, is its lower computational and memory requirements. Furthermore, 2D inputs accommodate using pre-trained nets, either directly or via transfer learning. However, an important drawback of such an approach is that anatomic context in directions orthogonal to the 2D plane is completely discarded. As discussed recently in [26], considering 3D MRI data directly, instead of slice-by-slice, can improve the performance of a segmentation method.

To incorporate 3D contextual information, de Brebisson et al. used 2D CNNs on images from the three orthogonal planes [27]. The memory requirements of fully 3D networks were avoided by extracting large 2D patches from multiple image scales, and combining them with small single-scale 3D patches. All patches were assembled into eight parallel network pathways to achieve a high-quality segmentation of 134 brain regions from whole brain MRI. More recently, Shakeri et al. [17] proposed a CNN scheme based on 2D convolutions to segment a set of subcortical brain structures. In their work, the segmentation of the whole volume was first achieved by processing each 2D slice independently. Then, to impose volumetric homogeneity, they constructed a 3D conditional random field (CRF) using scores from the CNN as unary potentials in a multi-label energy minimization problem.

So far, 3D CNNs have been largely avoided due to the computational and memory requirements of running 3D convolutions during inference. However, the ability to fully exploit dense inference is an important advantage of 3D CNNs over 2D representations [28]. While standard CNN approaches predict the class probabilities of each pixel independently from its local patch, fully convolutional networks (FCNNs) [29] consider the network as a large non-linear filter whose output yields class probabilities. This accommodates images of arbitrary size, as in regular convolution filters, and provides much greater efficiency by avoiding redundant convolutions/pooling operations. Recently, 3D FCNNs yielded outstanding segmentation performances in the context of brain lesions [30, 31].

1.2. Contributions

This study investigates a 3D and fully convolutional neural network for subcortical brain structure segmentation in MRI. 3D architectures have been generally avoided due to their computational and memory requirements during inference and, to the best of our knowledge, this work is the first to examine 3D FCNNs for subcortical structure segmentation. We address the problem via small kernels, allowing deeper architectures. We further model both local and global context by embedding intermediate-layer outputs in the final prediction, which encourages consistency between features extracted at different scales and embeds fine-grained information directly in the segmentation process. This contrasts with previous architectures (e.g., [31]), where global context is modelled via separate pathways and low-resolution images. Our model is efficiently trained end-to-end on a graphics processing unit (GPU), in a single learning stage, exploiting the dense inference capabilities of FCNNs.

We performed comprehensive experiments over two publicly available data sets. First, we demonstrate a state-of-the-art performance on the ISBR dataset. Then, we report a large-scale multi-site evaluation over 1112 unregistered subject data sets acquired from 17 different sites (ABIDE data set), with ages ranging from 7 to 64 years, showing that our method is robust to various acquisition protocols, demographics and clinical factors. Our method yielded segmentations that are highly consistent with a standard atlas-based approach, while running in a fraction of the time needed by such methods, and avoiding the computationally expensive and error prone registration/normalization steps. This makes it convenient for massive multi-site neuroanatomical imaging studies. We believe our work is the first to assess subcortical structure segmentation on such large-scale and heterogeneous data.

2. Methods and materials

Initially designed for image recognition or classification tasks, CNNs have recently shown a great potential for the problem of semantic segmentation. In particular, FCNNs have achieved state-of-the-art results on various image segmentation tasks. In section 2.1, we give a general presentation of this specific architecture, which is at the core of the proposed segmentation method. One of the main differences of our network, with respect to other 3D CNN architectures, is the use of features extracted at intermediate layers. The advantages of employing features at multiple scales are explained in Section 2.2. Thereafter, Section 2.3 presents the pre- and post-processing steps performed by our method on the data and output segmentations. Finally, Section 2.4 focuses on the study design and experimental setup, providing information on the datasets used in the study, implementation details of the tested network architectures, and the metrics used to evaluate the performance of these architectures.

2.1. 3D FCNNs for semantic segmentation

Convolutional neural networks (CNNs) are supervised models that are trained end-to-end to learn a hierarchy of features representing different levels of abstraction. Architectures of this type are typically made up of multiple convolution, pooling and fully-connected layers, the parameters of which are learned using back-propagation. One of the main differences with respect to classical neural networks is that CNNs exploit local connectivity. Unlike in a typical neural net, units in hidden layers of a CNN are not connected to all units of the previous layer. Instead, they are only connected to a small number of units, corresponding to a spatially localized region. The advantages of local connectivity are two-fold. Firstly, it reduces the number of parameters in the net, thereby limiting memory/computational requirements and reducing the risk of overfitting. Secondly, by modelling specific regions of an image, hidden units are able to detect local patterns in a data-driven manner. A significant benefit of this is that these learned features are often more discriminative than hand-engineered ones, and can more easily be accommodated to different problems.

A CNN normally contains several convolutional layers, each one composed of many hidden units. These hidden units are connected, possibly through previous network layers, to specific regions of the image, called receptive fields. The result of applying a convolution across an input image is known as a feature map. At each convolutional layer, the number of output feature maps is determined by the number of convolution kernels in this layer. Denote as m_l the number of convolution kernels in layer l of the network, and let x_{l-1}^n be the 3D array corresponding to the n-th input of layer l. The k-th output feature map of layer l is then given by:

$$y_l^k \;=\; f\big(\sum_{n=1}^{m_{l-1}} W_i^{k,n} \otimes x_{l-1}^n + b_l^k\big), \tag{1}$$

where $W_i^{k,n}$ is a kernel convolved (represented by $\otimes$) with each of the previous layers, b_l^k is the bias, and f is a non-linear activation function. Image y_l^k can be viewed as the activation of neurons of the k-th feature map, located in the l-th layer. Feature maps from the last convolutional layer are usually concatenated before being fed in the fully-connected layers. Furthermore, pooling or sub-sampling layers are typically applied after individual convolutional layers to achieve a certain level of spatial invariance. By reducing the resolution of feature maps, these layers also limit the number of network parameters, thereby preventing overfitting. Among the different pooling operations, max pooling is often used for its robustness to small pattern shifts in the image [32]. Lastly, neurons in the last layer (i.e., the classification layer) are grouped into $m = C$ feature maps, where C denotes the number of classes. The output of the classification layer L is then converted into normalized probability values via a softmax function. The probability score of class $c \in \{1, \ldots, C\}$ is computed as follows:

$$p_c \;=\; \frac{\exp\left(y_L^c\right)}{\sum_{c'=1}^{C} \exp\left(y_L^{c'}\right)}. \tag{2}$$

In their original form, CNNs were designed for image recognition or classification tasks, not for pixel-level segmentation. Traditional architectures (such as AlexNet or GoogLeNet) require an input image of fixed size $w \times h$, and completely discard the spatial information in the top-most layers. These last layers are generally fully-connected and output a vector of class scores p_c. In contrast, semantic segmentation, which combines both problems of object localization and delineation, requires a spatial map of class scores $p_c(x, y)$. One of the main drawbacks of using such CNNs to perform semantic segmentation is the loss of spatial information and resolution, produced by repeated convolutions and pooling operations. Fully Convolutional Networks (FCNNs) mitigate this problem by treating the network as a single non-linear convolution [29], thus allowing its application to images of arbitrary size. Moreover, because the spatial map of class scores is obtained in a single step, FCNNs avoid the redundant convolution and pooling operations performed by traditional CNNs, making them computationally more efficient. As depicted in Figure 1, standard CNN architectures can readily be converted into FCNNs by interpreting the fully connected layers at the end of the network as a large set of $1 \times 1 \times 1$ convolutions. For additional details on adapting CNNs classifiers for dense prediction, we refer the reader to Section 3.1 of [29] and Section 2.1 of [31].

2.2. Combining features across multiple scales

In various computer vision problems, reasoning across multiple levels of abstraction (i.e., *scales*) has proven beneficial. In optical flow, for instance, coarse levels of the image pyramid are useful for establishing broad correspondences, but finer levels are also needed to get accurate measurements of pixel displacement [33]. In the case of CNNs, the sequence of layers encodes features representing increasing levels of abstraction: the first convolutional layer typically models simple edge or blob detectors, whereas convolutional layers directly before the fully-connected ones model larger-scale and more complex structures.

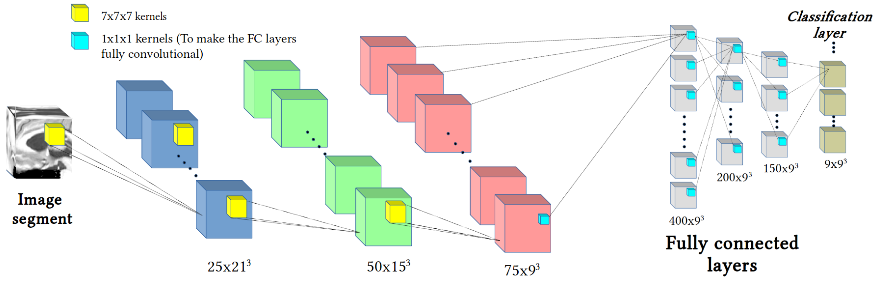


Figure 1: The baseline FCNN architecture (CNN_{base}), composed of 3 convolutional layers with kernels of size 7^3. This FCNN is obtained from a standard CNN by replacing the fully connected layers by a set of $1 \times 1 \times 1$ convolutional filters.

However, because pooling operations and convolution with non-zero skips decrease the resolution, these higher layers of the network encode coarser information (i.e., a greater receptive field but smaller resolution). To model both semantic information and local details, it is thus necessary to exploit features located in different layers of the network.

There are two main strategies to use multi-scale features in CNNs. In the first strategy, the input image is resampled at multiple scales of resolution, before being fed to the network [31]. These different inputs are then combined via separate pathways in the network. On the other hand, the second strategy uses a single resolution scale of the image as input (usually the highest available one), but combines the feature maps of various layers in the fully-connected layers [34, 35, 36, 37]. Although both strategies have led to significant improvements in performance, the later has two important advantages. First, because it has a single set convolution filters at each layer, instead of one per pathway, the features at different scales are more likely to be consistent with each other. Moreover, since features from intermediate layers are injected in the fully-connected layers, fine-grained information is used directly in the segmentation process. However, the drawback of this strategy is that it can lead to a large number of parameters in the fully-connected layers, which can make learning these parameters computationally complex. In the proposed method, we used the second strategy to include multi-scale information in the segmentation. Our experiment show that, even with a large number of parameters in the model, the learning process can be completed in a reasonable amount of time. Moreover, the high accuracy obtained by our model suggests that overfitting does not occur.

2.3. Pre- and post-processing steps

Data pre-processing steps are often required to ensure the performance of segmentation methods. Typical pre-processing steps for MRI data include the removal of non-brain tissues, like the skull, and bias field correction. For multi-subject or longitudinal studies, additional steps are often necessary to normalize intensities or align volumes across multiple scans. In [17], Shakeri et al. obtained a high accuracy for the subcortical parcellation of registered and normalized MRI volumes, showing their method to be superior to standard atlas-based approaches for this task. However, such elaborate data pre-processing has several disadvantages. First, aligning volumes to a template (e.g., MNI space) is a time-consuming operation, which would remove the computational benefit of using CNNs over atlas-based methods. Furthermore, training the network using data with a very specific and strict pre-processing reduces the network's ability of generalizing to unprocessed data, or data pre-processed differently.

In order to make our method robust to the different possible imaging protocols and parameters, we thus use a very simple pre-processing step, which includes volume-wise intensity normalization, bias field correction and skull-stripping. The first two transforms, both computationally inexpensive, are used to reduce the sensitivity of the network to contrast and intensity bias. Skull-stripping, although more time-consuming, can be performed without registration (e.g., see [38]). This step is used mostly to reduce the size of the input image by discarding non-interesting areas, and thus unnecessary computations.

Although the segmentations obtained using our network are generally smooth and close to manual labels, small isolated regions can sometimes appear in the segmentation. As post-processing step, we remove these small regions by keeping only the largest connected component from each class. Note that standard regularization approaches like CRFs [39] have also been tested, but did not lead to significant improvements in accuracy.

2.4. Study design and experiment setup

We start by describing the data used in our experiments, and then provide important details on the implementation of the proposed architecture.

2.4.1. Datasets

Two public datasets were used in the proposed experiments. To demonstrate the effectiveness our 3D FCNN method against other state-of-the-art semantic segmentation approaches, we first used the IBSR benchmark dataset, which contains pre-processed MRI data with corresponding ground-truth segmentations. Then, to validate its robustness with respect to various acquisition protocols, as well as demographics and clinical factors, we tested our method on the large-scale ABIDE dataset, containing the data of 1112 subjects acquired from 17 different sites. Details about these datasets are given below.

IBSR. A set of 18 T1-weighted MRI scans from the Internet Brain Segmentation Repository (IBSR) was employed to obtain quantitative measures of performance and compare our proposed method against competing approaches. These images were acquired at the Massachusetts General Hospital and are freely available at `http://www.cma.mgh.harvard.edu/ibsr/data.html`. In addition, the dataset also contains expert-labelled segmentations of 45 brain structures. Among these, a subset of 8 important subcortical structures were considered in this work: left and right thalamus, caudate, putamen, and pallidum. These structures were used in recent studies on brain parcellation (e.g., see [17]). All volumes have a size of 256 × 256 × 128 voxels, with voxel sizes ranging from $0.8 \times 0.8 \times 1.5$ mm^3 to $1.0 \times 1.0 \times 1.5$ mm^3. To get unbiased estimates of performance, and following the validation methodolody of [17], we employed a 6-fold cross validation strategy, where each fold is composed of 12 training examples (i.e., subjects), 3 validation examples and 3 test examples.

ABIDE. The Autism Brain Imaging Data Exchange (ABIDE) [40] was used as a second dataset in our experiments. ABIDE I involved 17 international sites, sharing previously collected resting state functional magnetic resonance imaging (R-fMRI), anatomical and phenotypic datasets made available for data sharing with the broader scientific community. This effort yielded a huge dataset containing 1112 subjects, including 539 from individuals with autism spectrum disorder (ASD) and 573 from typical controls (ages 7-64 years, median 14.7 years across groups). Some characteristics for each site are presented in Table 1. In the present study, we evaluated the segmentation of target subcortical brain structures by training and testing with data from different sites or age/diagnosis groups. For training, we considered 10 control subjects from 15 sites (indicated by an asterisk in the table), giving a total of 150 training examples. For validation, we used a single subject per site, leading to a validation set composed of 15 examples. Segmentation performance was evaluated on remaining subjects from all sites.

Unlike IBSR, the ABIDE dataset does not contain ground-truth segmentations of subcortical structures. Instead, we have used automatic segmentations obtained using the *recon-all* pipeline of the *FreeSurfer* 5.1 tool [41], which are freely available at `http://fcon_1000.projects.nitrc.org/indi/abide/`. This pipeline involves the following steps: motion correction, intensity normalization, affine registration of volumes to the MNI305 atlas, skull-stripping, nonlinear registration using the Gaussian Classifier Atlas (GCA), and brain parcellation. The outputs of this pipeline used in our study are the skull stripped, intensity normalized brain volumes in the unregistered subject space (i.e., *brain.mgz* files) and the sub-cortical labelling of these volumes (i.e., *aseg.mgz* files). For this dataset, the objectives of our experiment was to measure the impact of different imaging, demographic and clinical factors on the reliability of the proposed method. Another goal was to verify that our method could obtain segmentations similar to those of atlas-based approaches (e.g., the segmentation approach of FreeSurfer), but in a fraction of the time.

We exploit the broad age range of controls and ASD subjects in the ABIDE dataset to measure the impact of age on our method's performance. Note that brain volumes may differ between control subjects and individuals with autism. For example, Aylward et al. [42] found that brain development in autism follows an abnormal pattern, with accelerated growth in early life that results in brain enlargement in childhood. Nevertheless, brain volume in adolescents and adults with autism appears to be normal. Based on their findings, and following their age-wise grouping, we first divided subjects into three non-overlapping groups: <13 years, 13 to 18 years, and >18 years. Furthermore, various studies have

Site	**Number images**	**Voxel size (mm^3)**	**Control vs. ASD**	**Ages (Years)**
California Institute of Technology*	38	1.0×1.0×1.0	19/19	17.0-56.2
Carnegie Mellon University*	27	1.0×1.0×1.0	13/14	19-40
Kennedy Krieger Institute*	55	1.0×1.0×1.0	33/22	8.0-12.8
Ludwig Maximilians University Munich*	57	1.0×1.0×1.0	33/24	7-58
NYU Langone Medical Center*	184	1.3×1.0×1.3	105/79	6.5-39.1
Olin, Institute of Living at Hartford Hospital*	36	1.0×1.0×1.0	16/20	10-24
Oregon Health and Science University	28	1.0×1.0×1.1	15/13	8.0-15.2
San Diego State University*	36	1.0×1.0×1.1	22/14	8.7-17.2
Social Brain Lab BCN NIC UMC Groningen*	30	1.0×1.0×1.1	15/15	20-64
Stanford University*	40	0.86×1.5×0.86	20/20	7.5-12.9
Trinity Centre for Health Sciences	49	1.0×1.0×1.0	25/24	12.0-25.9
University of California, Los Angeles: Sample 1*	82	1.0×1.0×1.2	33/49	8.4-17.9
University of California, Los Angeles: Sample 2	27	1.0×1.0×1.2	14/13	9.8-16.5
University of Leuven: Sample 1*	29	0.98×0.98×1.2	15/14	18-32
University of Leuven: Sample 2*	35	0.98×0.98×1.2	20/15	12.1-16.9
University of Michigan: Sample 1	110	-×-×1.2	55/55	8.2-19.2
University of Michigan: Sample 2*	35	-×-×1.2	22/13	12.8-28.8
University of Pittsburgh School of Medicine*	57	1.1×1.1×1.1	27/30	9.3-35.2
University of Utah School of Medicine*	101	1.0×1.0×1.2	43/58	8.8-50.2
Yale Child Study Center	56	1.0×1.0×1.0	28/28	7.0-17.8

Table 1: Scan parameters and characteristics of sites included in the ABIDE dataset. An asterisk beside the site name indicates that data from this site were used for training.

identified physiological differences between ASD and healthy subjects in key brain regions, including the putamen [43], cerebellum [44], hippocampus [45], amygdala [46] and corpus callosum [47]. To account for these potential structural differences, we further split each age group into two sub-groups, containing control and ASD subjects respectively. Lastly, to evaluate the robustness of our method in unseen cohorts, the resulting subject groups were again split based on whether the subject is from a site used in training or not. Note that, in the case of subjects from sites used in training, only subjects from the test set are considered (i.e., no training example is used while measure the segmentation performance). A summary of the configuration and ID of each group is presented in Table 2.

	Site used in training						**Site NOT used in training**					
DX group	Control			ASD			Control			ASD		
Age group	<13	13-18	>18	<13	13-18	>18	<13	13-18	>18	<13	13-18	>18
Group ID	A	B	C	D	E	F	G	H	I	J	K	L

Table 2: Configuration of subject groups used in the proposed experiments.

2.4.2. Proposed architecture: implementation details

One of the main design choices, when using CNNs, is the network architecture. Selecting of the number of convolutional layers and the number of filters in these layers is a complex and problem-specific task. Likewise, the choice of pooling and activation functions can also have a significant impact on the networks performance and computational efficiency.

In this study, we investigate three different FCNN architectures. The first architecture, called CNN_{base}, is composed of 3 convolutional layers with 25, 50 and 75 feature maps, respectively, and a kernel size of 7^3. Each of these convolutional layer is followed by a Parametric Rectified Linear Unit (PReLU) layer [48], serving as activation function on the output features. Furthermore, three fully-connected layers are added after the last convolutional layer to model the relationship between features and class labels. These PReLU and fully-connected layers, which are common to all three tested architectures, are further described below. The CNN_{base} architecture, depicted in Figure 1, is employed as a baseline to generate “standard” or “control” segmentations.

In the second architecture, denoted as $\text{CNN}_{\text{single}}$, each convolutional layer is replaced by three successive convolutional layers with the same number of kernels, but smaller kernel sizes: 3^3 instead of 7^3. By using these smaller kernels, we obtain a deeper architecture while having fewer parameters in the network. Consequently, the network can learn a more complex hierarchy of features, with a reduced risk of overfitting. This fact is supported by the findings reported in [49] for 2D CNNs, and in [31] for 3D CNNs.

The third architecture, which we call $\text{CNN}_{\text{multi}}$, corresponds to the proposed 3D FCNN of this work. This architecture extends the $\text{CNN}_{\text{single}}$ model by exploiting features at multiple scales in the network. Specifically, the outputs of PReLU-activated convolutional layers 3, 6 and 9 are concatenated to form the input of the first fully-connected layer. This multi-scale architecture is illustrated in Figure 2, where the blue and green arrows correspond to the injection of convolutional layers 3 and 6, respectively, into the fully-connected layers (grey-colored in the figure).

For all three architectures, the fully-connected layers were composed of 400, 200 and 150 hidden units, respectively. These layers are followed by a final classification layer, which outputs the probability maps for each of the 9 classes: 8 for each of the subcortical structures (left and right) and one for the background. The $\text{CNN}_{\text{multi}}$ architecture proposed in this paper is thus composed of 13 layers in total, with the following layout: 9 convolutional layers (each followed by a PReLU layer), 3 fully-connected layers, and the classification layer.

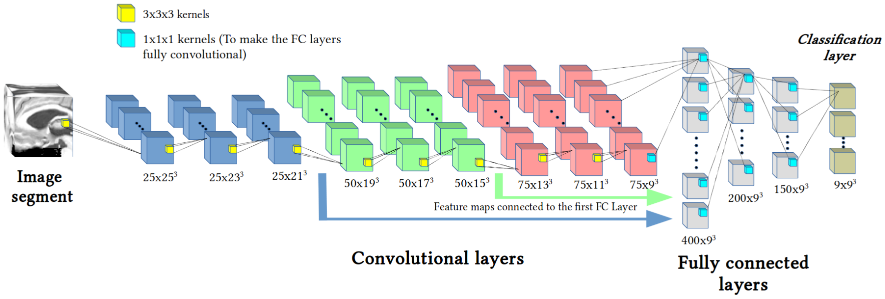


Figure 2: The proposed $\text{CNN}_{\text{multi}}$ architecture. Compared to CNN_{base}, this architecture achieves a deeper network by using smaller convolution kernels (3^3). Unlike $\text{CNN}_{\text{single}}$, it also uses the feature maps of intermediate layers directly in the fully-connected layers, as represented by the arrows.

Furthermore, the number of kernels in each convolutional layer (from first to last) is as follows: 25, 25, 25, 50, 50, 50, 75, 75 and 75.

For the activation function after each convolutional layer, we used the Parametric Rectified Linear Unit (PReLU) [48] instead of the popular Rectified Linear Unit (ReLU). This function can be formulated as

$$f(x_i) \ = \ \max(0, x_i) \ + \ a_i \cdot \min(0, x_i), \tag{3}$$

where x_i defines the input signal, $f(x_i)$ represents the output, and a_i is a scaling coefficient for when x_i is negative. While ReLU employs predefined values for a_i (typically equal to 0), PReLU requires learning this coefficient. Thus, this activation function can adapt the rectifiers to their inputs, improving the network's accuracy at a negligible extra computational cost. Note that, unlike traditional CNNs, the three tested architectures did not include any pooling layer. This is because pooling operations down-sample feature maps in the network (typically by a scale of two, for each pooling layer), which would lead to low-resolution maps in the last layers. By avoiding down-sampling features, we thus have a more accurate segmentation.

As mentioned above, using smaller kernels allows increasing the depth of the network. Having a deeper network, however, makes the training process more difficult. Compared to traditional sigmoid or hyperbolic tangent activation functions, network with rectifier units typically converge faster by avoiding the vanishing gradients problem [19]. Yet, a bad initialization can still hamper the training of a highly non-linear model. Initialization of deep CNNs has been mostly performed by assigning random normal-distributed values to kernel weights. As demonstrated in [49], initializing weights with fixed standard deviations may lead to poor convergence. To overcome these limitations, we adopted the strategy proposed in [48], and used in [31] for segmentation, that allows extremely deep architectures (e.g., 30 convolutional or fully-connected layers) to converge rapidly. In this strategy, weights in layer l are initialized based on a zero-mean Gaussian distribution of standard deviation $\sqrt{2/n_l}$, where n_l denotes the number of connections to units in that layer. For example, in the first convolutional layer of Figure 2 (*bottom*), the input is composed of single-channel (i.e., grey level) image segments and kernels have a size of 3^3, thus the standard deviation is equal to $\sqrt{2/(1 \times 3 \times 3 \times 3)} = 0.2722$. On the other hand, the second layer takes as input 25 features maps from the previous layer, therefore the standard deviation would be equal to $\sqrt{2/(25 \times 3 \times 3 \times 3)} = 0.0544$.

The optimization of network parameters is performed with stochastic gradient descent (SGD), using cross-entropy as cost function. However, since our network employs 3D convolutions, and due to the large sizes of MRI volumes, dense training cannot be applied to whole volumes. Instead, volumes are split into B smaller segments, which allows dense inference in our hardware setting. Let θ be the network parameters (i.e., convolution weights and biases), and denote as $\mathcal{L}$ the set of ground-truth labels such that $L_s^v \in \mathcal{L}$ is the label of voxel v in the s-th image segment. Following [31], we defined the cost function as

$$J(\theta; \mathcal{L}) \;=\; -\frac{1}{B \cdot V} \sum_{s=1}^{B} \sum_{v=1}^{V} \log\left(p_{L_s^v}(X_v)\right), \tag{4}$$

where $p_c(X_v)$ is the output of the classification layer for voxel v and class c. In [31], Kamnitsas et al. found that increasing the size of input segments in training leads to a higher performance, but this performance increase stops beyond segment sizes of 25^3. In their network, using this segment size for training, score maps at the classification stage were of size 9^3. Since our architecture is one layer deeper, and to keep the same score map sizes, we set the segment size in our network to 27^3.

Our 3D FCNN was initially trained for 50 epochs, each one composed of 20 subepochs. At each subepoch, a total of 500 samples were randomly selected from the training image segments, and processed in batches of size 5. However, we observed that the performance of the trained network on the validation set did not improve after 30 epochs, allowing us to terminate the training process at this point. As other important meta-parameters, the training momentum was set to 0.6 and the initial learning rate to 0.001, being reduced by a factor of 2 after every 3 epochs.

To implement our network, we adapted the 3D FCNN architecture of Kamnitsas et al. [31]. Their architecture was developed using Theano, a CPU and GPU mathematical compiler for implementing deep learning models [50]. The PC used for training is an Intel(R) Core(TM) i7-6700K 4.0GHz CPU, equipped with a NVIDIA GeForce GTX 960 GPU with 2 GB of memory. Training our network took a little over 2 hours per epoch, and around 2 days and a half for the fully trained CNN.

2.4.3. Evaluation

Various comparison metrics have been used in the literature to evaluate the accuracy of segmentation methods. Since these metrics yield different information, their choice is important and must be considered in the appropriate context. Although volume-based metrics, such as Dice similarity coefficient (DSC) [51], have been broadly used to compare segmentation results, they are fairly insensitive to the precise contour of segmented regions, which only has small impact on the overall volume. However, two segmentations with a high spatial overlap may exhibit clinically relevant differences in their boundaries. To measure such differences, distance-based metrics like the Modified Hausdorff distance (MHD) are typically used.

Dice similarity coefficient. Let V_{ref} and V_{auto} denote the binary reference segmentation and the automatic segmentation, respectively, of a given tissue class for a given subject. The DSC is then defined as

$$\mathrm{DSC}\big(V_{\mathrm{ref}}, V_{\mathrm{auto}}\big) \;=\; \frac{2 \mid V_{\mathrm{ref}} \cap V_{\mathrm{auto}} \mid}{\mid V_{\mathrm{ref}} \mid + \mid V_{\mathrm{auto}} \mid} \tag{5}$$

DSC values are comprised in the $[0,1]$ range, where 1 indicates perfect overlapping and 0 represents no overlapping at all.

Modified Hausdorff distance. Let P_{ref} and P_{auto} denote the sets of voxels within the reference segmentation and the automatic one, respectively. The MHD can be then defined as

$$\mathrm{MHD}\big(P_{\mathrm{ref}}, P_{\mathrm{auto}}\big) \;=\; \max\Big\{d(P_{\mathrm{ref}}, P_{\mathrm{auto}}), d(P_{\mathrm{auto}}, P_{\mathrm{ref}})\Big\}, \tag{6}$$

where $d(P, P')$ is the maximum distance between a voxel in P and its nearest voxel in P'. In this case, smaller values indicate higher proximity between two point sets, and thus a better segmentation.

3. Results

We first evaluate our architecture on the IBSR dataset, which contains ground-truth segmentations of various brain structures. This dataset has been used in numerous studies to evaluate subcortical parcellation methods. The next subsection reports the results of this evaluation. Furthermore, to better understand the impact of a deeper network with smaller kernels and of using multi-scale features, we report results obtained by the baseline architecture (i.e., $\mathrm{CNN}_{\mathrm{base}}$), and by the deeper network with (i.e., $\mathrm{CNN}_{\mathrm{multi}}$) and without (i.e., $\mathrm{CNN}_{\mathrm{single}}$) features extracted at intermediate levels. The results of this analysis are presented in Section 3.2. For notation simplicity, we now on denote brain structures by their first two characters, indicating within parenthesis their location on brain, i.e left or right hemisphere. For example, the caudate structure in the left brain side will be referred to as Ca(L).

3.1. IBSR dataset

Figure 3 depicts the results of our architecture, with the DSC and MHD values plotted for the left- and right-side structures. We see that the segmentation of the pallidum, both left and right, was significantly less accurate than other structures (i.e., thalamus, caudate and putamen), likely due to the smaller size of this brain structure. Furthermore, we observe that the segmentation of all four subcortical structures is slightly more accurate in the right hemisphere, although the differences are not statistically significant following a Wilcoxon signed-rank test.

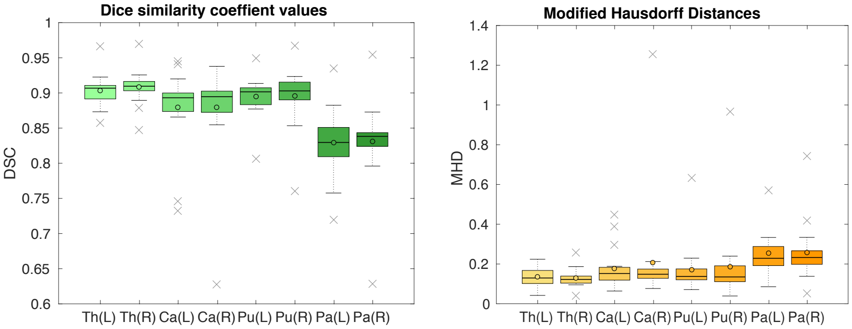


Figure 3: Segmentation accuracy obtained by the $\text{CNN}_{\text{multi}}$ architecture, for each brain structures, on subjects of the IBSR dataset.

In light of the various studies using the IBSR dataset as segmentation benchmark, the results obtained by our method are state-of-the-art (Table 4). Specifically, when comparing against the recent work of Shakeri et al. [17], which also proposed FCNNs and used the same validation methodology, our method achieved an DSC improvement ranging from 5% (in the thalamus) to 13% (in the caudate).

3.2. ABIDE dataset

To evaluate the impact of having a deeper network and multi-scale features, all examples of the ABIDE dataset were segmented using the three tested CNN configurations (i.e., CNN_{base}, $\text{CNN}_{\text{single}}$ and $\text{CNN}_{\text{multi}}$). The mean DSC and MHD of these segmentations are presented in Table 3. Recall that these accuracy measures were computed using the labels obtained from FreeSurfer, since ground-truth segmentations were not available. The two main objectives of this experiment are: (i) to show that our results are consistent with those obtained by a standard atlas-based approach, but with the advantage that our method runs in a fraction of the time needed by such approach; and (ii) to compare results across different subject groups and acquisition sites.

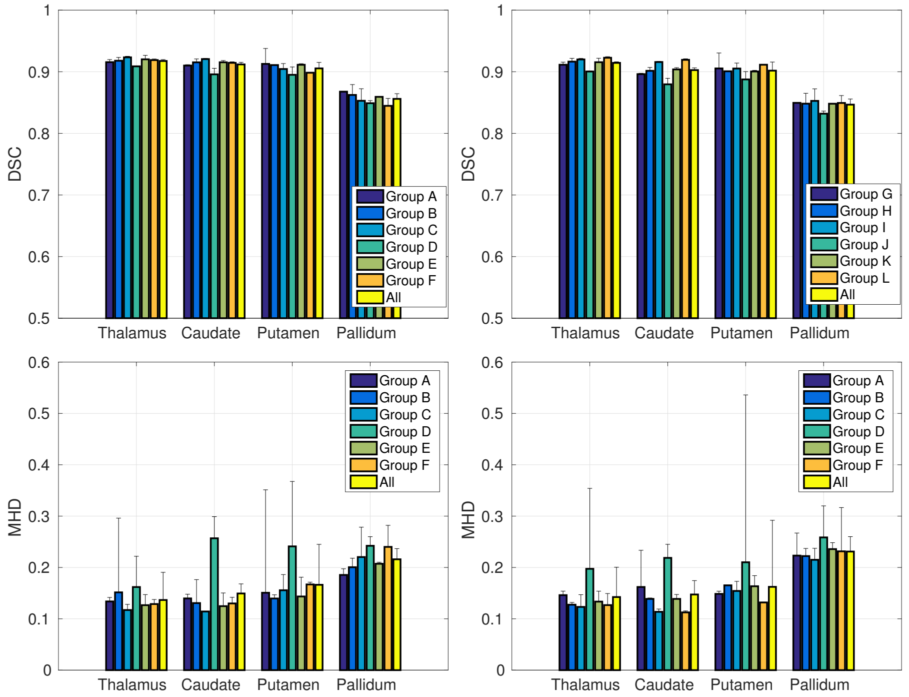
We first observe that having a deeper network, via smaller kernels, increases the segmentation performance in both metrics. In a one-sided Wilcoxon signed-rank test, the mean DSC and MHD of $\text{CNN}_{\text{single}}$ is statistically better (i.e., higher for DSC and lower for MHD) than CNN_{base}, with $p < 0.01$. Likewise, when features extracted at intermediate layers are fed into the first fully connected layer, the proposed $\text{CNN}_{\text{multi}}$ network generated more reliable segmentations, both in terms DSC and MHD. These results are also statistically significant, with $p < 0.01$, in a Wilcoxon signed-rank test.

	Structure	$\mathbf{CNN_{base}}$	$\mathbf{CNN_{single}}$	$\mathbf{CNN_{multi}}$
Mean DSC	Thalamus	0.8987 (0.002)	0.9039 (0.0052)	**0.9156** (0.0012)
	Caudate	0.8979 (0.0012)	0.9011 (0.0002)	**0.9073** (0.0021)
	Putamen	0.8909 (0.0017)	0.8992 (0.0009)	**0.9041** (0.0014)
	Pallidum	0.8381 (0.0096)	0.8497 (0.0096)	**0.8621** (0.0095)
Mean MHD	Thalamus	0.1487 (0.0087)	0.1462 (0.0153)	**0.1405** (0.0002)
	Caudate	0.1710 (0.0154)	0.1583 (0.0213)	**0.1557** (0.0165)
	Putamen	0.2074 (0.0296)	0.1742 (0.0467)	**0.1706** (0.0296)
	Pallidum	0.2487 (0.0205)	0.2305 (0.0208)	**0.2232** (0.0232)

Table 3: Mean DSC and MHD (standard deviation between brackets), obtained by the three tested CNN architectures on the ABIDE dataset. Bold font numbers indicate the best result among the three architectures.

Figure 4 plots the mean DSC and MHD values obtained by our $\text{CNN}_{\text{multi}}$ architecture for each of the subject groups in Table 2. These values are grouped by subcortical structure of interest, i.e., thalamus, caudate, putamen and pallidum. For each structure, an additional bar is added, giving the mean DSC and MHD obtained on subjects of all groups together. Across all subject groups, the segmentations produced by our CNN achieved mean DSC values above 0.90 for all structures except the pallidum, which had a mean DSC of 0.85. Likewise, mean MHD values were below 0.25 mm in all subject groups and for all four subcortical structures. These results are consistent with those obtained for the IBSR dataset.

Analyzing the results obtained using data from sites considered in training (groups A-F), we observe that mean DSC values obtained for control subjects (groups A-C) are usually higher than for ASD subjects (groups D-E). For instance, putamen segmentation in control subjects less than 13 years old yielded a mean DSC of 0.9127, compared to 0.9055 for ASD subjects in the same age group. The same trend is seen for distance similarities, for example in the caudate, where a mean MHD of 0.1397 was obtained for control subjects, versus 0.2568 for ASD subjects. These results illustrate that physiological differences



a) Data employed in training b) Unseen data

Figure 4: Mean DSC and MHD values obtained for subject data used during training, and for unseen data.

related to autism, especially in young subjects, can have a small impact on segmentation accuracy. Looking at the impact of subject age on results, it can be seen that the segmentation of the thalamus and caudate improves as the subject gets older, in both control and ASD subjects. The relationship between subject age and segmentation accuracy in these structures is further illustrated in Figure 5, which gives the scatter plot of DSC versus age in the left/right thalamus and caudate, considering all control and ASD subjects together. In each plot, the Spearman rank correlation coefficient and corresponding p-value are given as variables r and p. Note that p-values have been corrected using the Bonferroni procedure, to account for the multiple comparisons (8 structures). We notice a weak but statistically significant correlation, with $p < 0.01$, validating our previous observation. It is also worth noting a greater variance in accuracy occurring for younger subjects, most of the low accuracy values observed for ages less than 20 years old. This is consistent with the fact that the brain is continuously developing until adulthood, and suggests that the physiological variability of younger subjects may not be completely captured while training the CNN.

The same patterns can be observed when segmenting subjects from sites not used in training (groups G-L). Particularly noticeable is the relationship between age and accuracy, which can be seen in all structures, and in both control and ASD subjects. Comparing with results obtained on data from sites used in training, we find no statistically significant difference in accuracy (DSC or MHD), for any brain structure. This suggests that the proposed method can generalize to acquisition protocols and imaging parameters not seen in training. Overall the results of these experiments illustrate that our method is robust to various clinical, demographics and site-related factors.

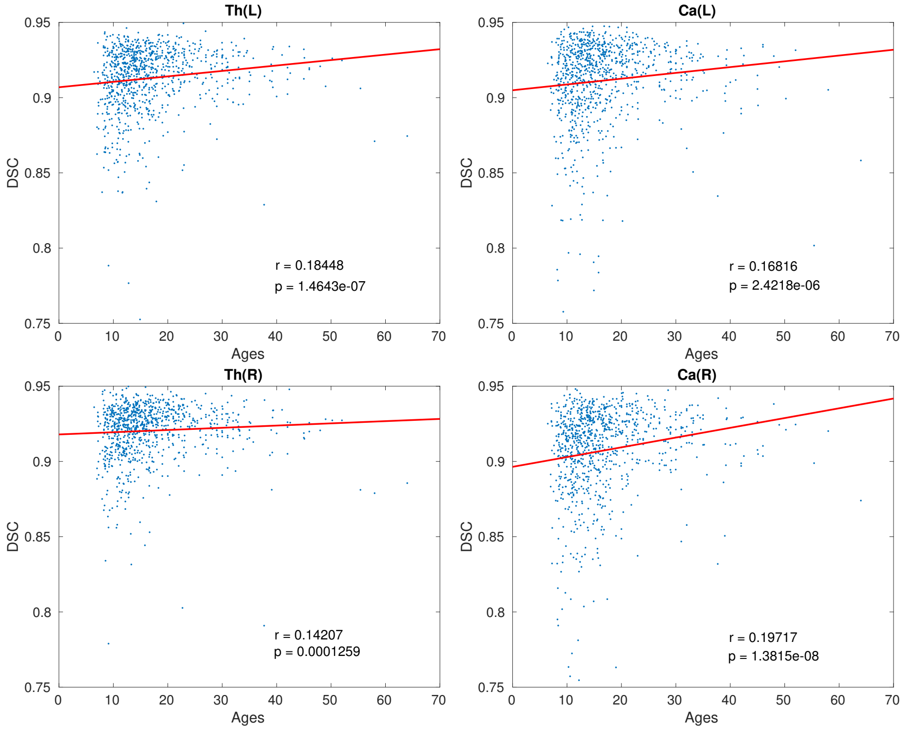


Figure 5: Scatter plots of left and right thalamus and caudate segmentation performance regarding DSC and subject age.

Figures 6 and 7 give visual examples of segmentations obtained by our 3D FCNN architecture and standard references contoured by *FreeSurfer*. In these figures, the first and third columns display the reference *FreeSurfer* contours, whereas the second and fourth columns contain segmentations of the corresponding images by our method. It can be observed that the segmentations generated by our proposed architecture are significantly smoother than those of *FreeSurfer*, regardless of the subject group (i.e diagnosis, age, site employed or not in training). We also notice that our system is better at identifying thin regions in the structures of interest, for instance, the lower extremities of pallidum (green regions).

To better understand the features learned automatically by the proposed

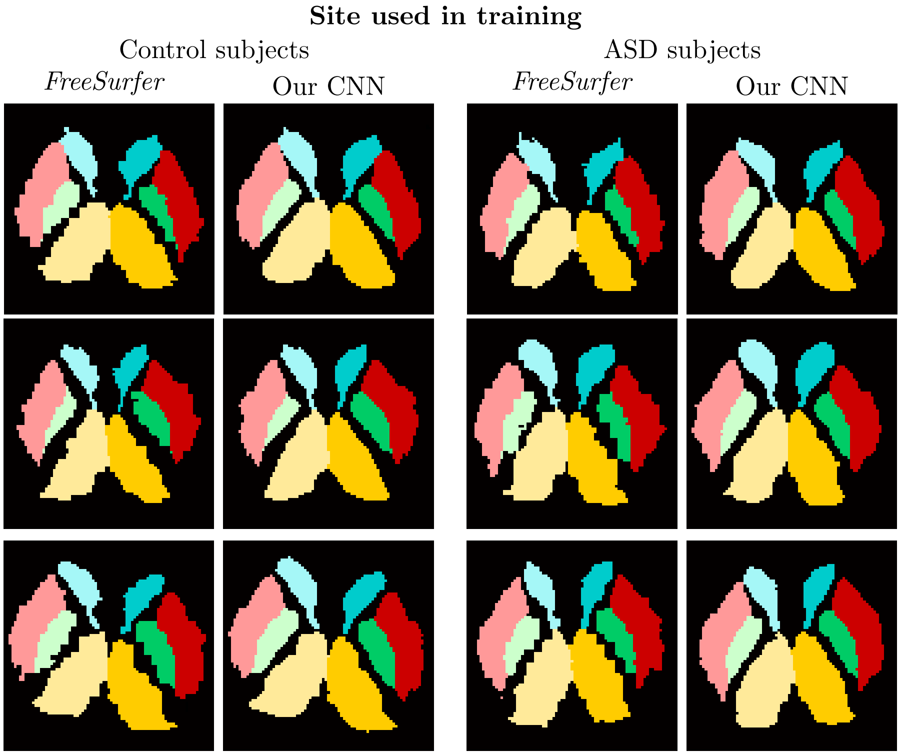


Figure 6: Visual examples of our 3D-FCNN architecture compared with the standard references contoured by *FreeSurfer*, for six test subjects from sites used in training.

CNN, Figure 8 shows examples of feature map activations obtained for a given input patch (cyan box in the figure). Each column corresponds to a different CNN layer, left-side columns corresponding to shallow layers, and right-side columns to deep layers in the network. Likewise, images in each row correspond to a randomly selected activation of the layer's feature map. Although difficult to analyze, we notice that activation values in initial layers mainly indicate the presence of strong edges or boundaries, whereas those in deeper layers of the network represent more complex structures. In particular, images in the last two columns (i.e., convolutional layers of the network) roughly delineate the right caudate. Note that 2D images are used here for visualization purposes and that both input patches and features map activations are actually in 3D.

As previously explained, score maps (i.e., class probabilities, ranging from 0 to 1) are obtained at the end of the network, before the voxels are assigned to the target labels. To illustrate this output, Figure 9 shows an example of probability maps for a given slice of the volume. Red pixels indicate probability values close to 1, and blue pixels near 0. Each image of the figure gives the probability map of a specific structure of interest, including the background. It can be seen that generated probability maps are well defined, reflecting the actual

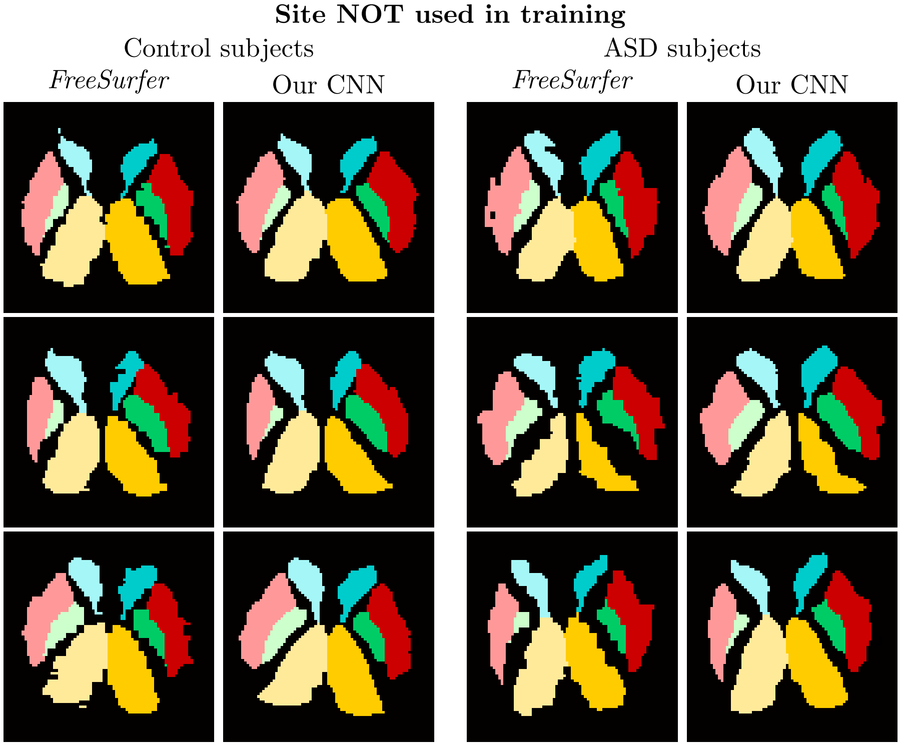


Figure 7: Visual examples of our 3D-FCNN architecture compared with the references standard contoured by *FreeSurfer* for six test subjects from sites not used in training.

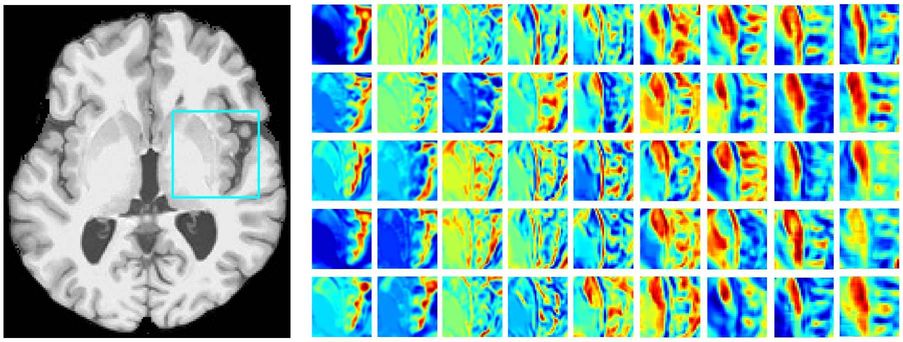


Figure 8: Feature map activations in all convolutional layers of the CNN (*right*), obtained for a given patch of the input MRI image (*left*). Each column corresponds to a different convolutional layer, from shallow to deeper, and each image in a row to a features map activation randomly selected in the layer.

contours of the imaged structures (first subfigure of the set). This suggests that these probability maps can be used directly for segmentation, without requiring additional, and potential computationally expensive, spatial regularization.

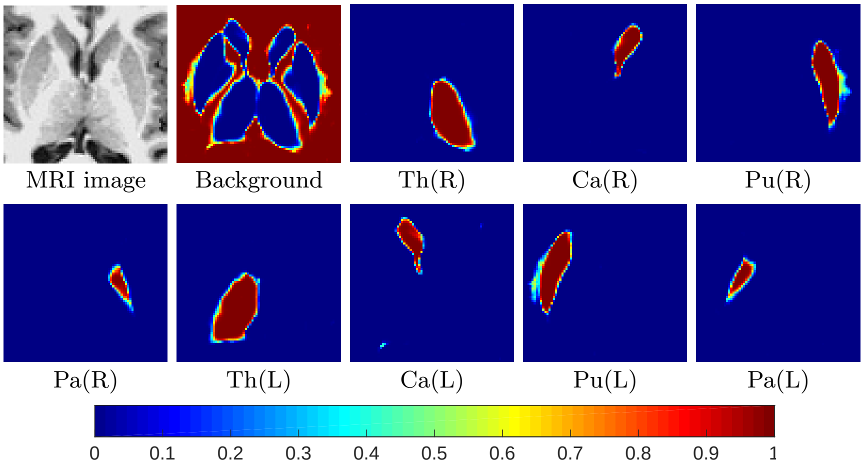


Figure 9: Probability maps generated by the proposed 3D FCNN for the background and the eight structures given an input MRI image. Note that input MRI has been cropped for better resolution.

Figure 10 displays a smoothed version of the 3D output provided by our CNN architecture. These images, which were rendered using the Medical Interaction ToolKit (MITK) software package [52], highlight the spatial consistency of the obtained segmentation.

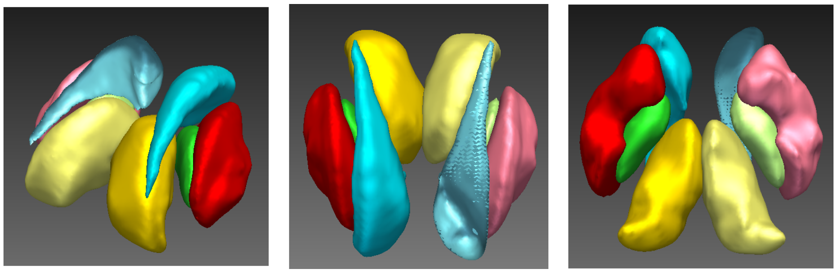


Figure 10: Different views of a smoothed version of contours provided by our automatic segmentation system. In these images, the thalamus, caudate, putamen and pallidum are respectively depicted in yellow, cyan, red and green.

Our method took on average 2-3 minutes for segmenting one test subject from the multi-site ABIDE dataset (nearly two days for all 947 subjects). All automatic contours and probability maps generated by our deep network were made publicly available at: `https://github.com/josedolz/3D-F-CNN-BrainStruct`.

4. Discussion

We presented a 3D fully-convolutional neural network architecture for the semantic segmentation of subcortical brain structures in MRI. In contrast to existing CNN approaches for this task, which employ a 2D slice-by-slice strategy, the proposed architecture captures volumetric information directly with 3D convolutions. Moreover, by using smaller convolution kernels, we obtained a deeper network with less parameters. This strategy removed the need for pooling operations that down-sample feature maps, thereby yielding more accurate segmentations. Another innovative aspect of the proposed method is the use of multi-scale features, where the output of intermediate layers of the network are embedded directly into the fully-connected layers. Our FCNN architecture is efficiently trained end-to-end on a graphics processing unit (GPU) in a single learning stage for the voxel-wise labeling of MRI volumes.

We conducted a comprehensive quantitative evaluation of our method using two publicly available datasets. In a first set of experiments, the segmentation accuracy of our method was measured with respect to the ground-truth segmentations of the IBSR dataset, and compared to recently proposed methods for the task of brain parcellation. As reported in Table 4, our method obtained state-of-the-art performance on this dataset, with mean DSC values ranging from 0.83 to 0.91 and mean MHD values between 0.13 mm and 0.26 mm (Figure 3).

The ABIDE dataset was then used to demonstrate the reliability of the proposed method for large-scale data sets acquired at multiple sites. The impact of various factors, including age, diagnosis group (i.e., healthy control or ASD) and site was measured by dividing the data into different test groups. Considering all test subjects together, our method obtained segmentations consistent with those of *FreeSurfer*, with mean DSC between 0.86 and 0.92 and mean MHD ranging from 0.14 mm to 0.22 mm, across the target brain structures. The accuracy of the proposed architecture is statistically higher than two other CNN architectures, which do not use multi-scale features and small kernels (Table 3).

Considering the diagnosis group of subjects, segmentations obtained for both control and ASD subjects were of high quality, with similar mean DSC and MHD values (Figure 7). Since ASD subjects are likely to have morphological (e.g., volumetric) differences in brain regions like the putamen [43], hippocampus [45] or amygdala [46], compared to healthy sujects, this suggests that our method is robust to such differences. Analyzing the results according to subject age group, we noticed a slightly lower segmentation accuracy for younger subjects. This makes sense, considering the fact that the brain is continuously developing until adulthood, and that young subjects have a larger variability during their development process. However, it has been found that brain development in autism follows an abnormal pattern, with accelerated growth in early life, which results in brain enlargement during childhood [42]. Therefore, there may be some intermediate states of brain development in early ages of control and ASD subjects that were not fully captured by the CNN during training. Finally, by achieving a comparable performance on subjects from sites used in training and subject from other sites, we demonstrated that our method is robust to the various imaging parameters and protocols.

The automated segmentation of brain regions in MRI is a challenging task due to the structural variability across individuals. To tackle this problem, a considerable amount of approaches have been proposed during the last decade (Table 4), many of which are based on atlases. Although atlas-based segmentation has been used successfully for subcortical brain structure segmentation, a single atlas is often unsuitable for capturing the full structural variability of subjects in a given neuroimaging study. Several strategies have been presented to overcome the limitation of single atlas segmentation, for instance using multiple atlases alongside label fusion techniques [53]. Nevertheless, one of the main drawbacks shared by all atlas-based methods is their dependency to the image registration step, which is both time-consuming and prone to errors. Some recent studies have reported segmentation times per subject in the range of 10 to 15 hours when employing *FreeSurfer* [54, 55]. In [15], Powell et al. presented an approach based on artificial neural networks as an alternative to atlas-based methods. However, registration was also a key component of their segmentation scheme, thus having the same drawbacks as atlas-based techniques. Also using machine learning, a CNN was proposed in [17] for the task of subcortical brain parcellation. Although the registration of subjects volumes was not initially required, the authors tested their CNN on data pre-registered to the Talairach space. As demonstrated by our experiments, our approach has the advantage of being alignment independent. This property is of great importance, particularly from the clinical point of view, since scans from different subjects are not typically aligned. Moreover, since non-linear registration is a computationally expensive process, avoiding it can significantly reduce segmentation times.

Although 2D CNNs have led to record-breaking performances in various computer vision tasks, their usefulness for 3D medical images is more limited. Numerous strategies have been proposed to mitigate this, for instance, considering all three orthogonal planes [27], or using single slices with a regularization scheme (e.g., CRF) to impose volumetric homogeneity [17]. Although these techniques have helped improving segmentation results, they lack the ability to capture the full spatial context of 3D images. By using 3D convolutions, our approach can better capture spatial context in volumetric data. This is reflected by a performance improvement with respect to typical 2D CNN models. Another noteworthy point is the ability of our method to successfully segment subjects from sites that were not employed during training. Differences in scanners or acquisition protocols, for instance, can introduce a significant bias on the appearance of images (e.g., alignment, contrast, etc.), and the heterogeneity of multi-site data has been a stumbling block for large-scale neuroimaging studies. As confirmed by our results, incorporating training samples from different sites, which cover a wider range of variability, allowed us to alleviate this problem.

For the experiments on the ABIDE dataset, the reference contours used for training our CNN were obtained with *Freesurfer*, which is considered as a standard approach to subcortical brain labeling. While expert-labelled contours would have provided a more reliable validation of our approach, it was found that the contours obtained by our method were consistent with those of *Freesurfer*.

Work	Method	Structures	DSC	Dataset
-	Majority voting	Thalamus Caudate Putamen Pallidum	0.83 0.69 0.74 0.75	IBSR
Heckemann et al. [56] (2006)	Atlas	Thalamus Caudate Putamen Pallidum	0.90 0.90 0.90 0.80	Own dataset
Han et al. [57] (2007)	Atlas	Thalamus Caudate Putamen Pallidum	0.88 0.84 0.85 0.76	Own dataset
Linguraru et al. [58] (2007)	Atlas	Thalamus Caudate Putamen Pallidum	0.88 0.82 0.86 0.79	IBSR
Bazin et al. [59] (2008)	Atlas	Thalamus Caudate Putamen Pallidum	0.77 0.78 0.82 -	IBSR
Powell et al. [15] (2008)	Artificial Neural Network	Thalamus Caudate Putamen Pallidum	0.88 0.84 0.85 -	Own dataset
Artaechevarria et al. [60] (2009)	Atlas	Thalamus Caudate Putamen Pallidum	0.88 0.83 0.87 0.81	IBSR
Ciofolo et al. [61] (2009)	Atlas	Thalamus Caudate Putamen Pallidum	0.77 0.60 0.66 0.56	IBSR
Lotjonen et al. [10] (2010)	Atlas	Thalamus Caudate Putamen Pallidum	0.89 0.85 0.90 0.80	IBSR
Sabuncu et al. [62] (2010)	Atlas	Thalamus Caudate Putamen Pallidum	0.91 0.87 0.89 0.84	Own dataset
Patenaude et al. [63] (2011)	Bayesian model	Thalamus Caudate Putamen Pallidum	0.89 0.83 0.88 0.79	Own dataset
Rousseau et al. [64] (2011)	Atlas	Thalamus Caudate Putamen Pallidum	0.88 0.87 0.87 0.64	IBSR
Asman et al. [65] (2014)	Atlas	Thalamus Caudate Putamen Pallidum	0.89 0.90 0.89 0.84	OASIS
Wang et al. [66] (2014)	Atlas	Thalamus Caudate Putamen Pallidum	0.89 0.75 0.88 0.84	OASIS
Shakeri et al. [17] (2016)	2D FCNN + CRF	Thalamus Caudate Putamen Pallidum	0.87 0.78 0.83 0.75	IBSR
Bao et al. [67] (2016)	2D CNN	Thalamus Caudate Putamen Pallidum	0.90 0.87 0.88 0.80	IBSR
Our CNN	3D FCNN	Thalamus Caudate Putamen Pallidum	**0.92** **0.91** **0.90** **0.86**	IBSR
Our CNN	3D FCNN	Thalamus Caudate Putamen Pallidum	**0.92** **0.92** **0.91** **0.86**	ABIDE

Table 4: Summary of brain subcortical structures segmentation methods. Majority voting method is employed with the IBSR dataset to demonstrate that the proposed approach actually learns from training data.

Furthermore, a visual inspection of the results revealed that our method's contours were more regular than those obtained by *Freesurfer*. This suggests our CNN to be a suitable alternative to *Freesurfer*'s parcellation pipeline. Nevertheless, an evaluation involving trained clinicians would be necessary to fully validate this assertion.

Analyzing the results, we observed that the segmentation of several subject data differed considerably from others. Upon visual inspection, we found that the corresponding MRI images had a poor quality (e.g., motion artifacts), and decided not to include them in the evaluation. Figure 11 shows examples of 2D slices (in axial view) of two subjects with problematic data.

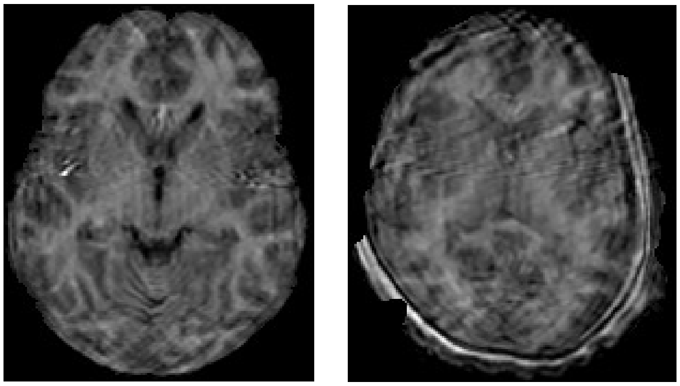


Figure 11: Axial slices from a bad quality scan of two subjects that were excluded from the evaluation.

Many modifications to the proposed architecture are possible. For example, the architecture could have a different number of convolutional/fully-connected layers, or a different number of filters/units in these layers. Several parameters settings were tested in preliminary experiments, to come up with a definitive architecture. Although the chosen parameters were found to perform well on the test data, they might not be optimal for other datasets. Despite this, small variations in the architecture are unlikely to have a large impact on performance. In future work, it would be interesting to further investigate the optimization of these parameters, such that they could be tuned automatically for a specific task and target data. In [31], Kamnitsas et al. found that different segment sizes as input to their network led to differences in performance. In our study, we used input sizes that worked well for their specific application, i.e. brain lesion segmentation. Although our target problem also uses brain images, characteristics of both problems are different, and the effect of input sizes on performance might also differ. We thus intend to investigate the impact of this factor in a subsequent study.

Another important aspect of CNNs is the transferability of knowledge embedded in the pre-trained CNNs, i.e transfer learning. The use of pre-trained CNNs has been already investigated in previous works. Nevertheless, available pre-trained models mainly come from 2D convolutions and its use is often tailored to the same application. We believe that pre-trained CNNs can be successfully used for different applications sharing the same nature, even if their objectives differ. For instance, our 3D FCNN trained on subcortical brain structures may be employed as pre-trained network to segment cardiac images.

5. Conclusion

We presented a method based on fully connected convolutional networks (FCNNs) for the automatic segmentation of subcortical brain regions. Our approach is the first to use 3D convolutional filters for this task. Moreover, by exploiting small convolution kernels, we obtained a deeper network that has fewer parameters and, thus, is less prone to overfitting. Local and global context were also modelled by injecting the outputs of intermediate layers in the network's fully connected layers, thereby encouraging consistency between features extracted at different scales, and embedding fine-grained information directly in the segmentation process.

We showed our multi-scale FCNN approach to obtain state-of-the-art performance on the well-known IBSR dataset. We then evaluated the impact of various factors, including acquisition site, age and diagnosis group, using 1112 unregistered subject data sets acquired from 17 different sites. This *large-scale* evaluation showed our method to be robust to these factors, achieving outstanding accuracy for all subjects groups. Additionally, these experiments have highlighted the computational advantages of our approach compared to atlas-based methods, by obtaining consistent segmentation results in a fraction of the time. In summary, we believe this work to be an important step toward the adoption of automatic segmentation methods in large-scale neuroimaging studies.

Acknowledgments

This work is supported by the National Science and Engineering Research Council of Canada (NSERC), discovery grant program, and by the ETS Research Chair on Artificial Intelligence in Medical Imaging.

References

References

1. T. van Erp, D. Hibar, J. Rasmussen, D. Glahn, G. Pearlson, O. Andreassen, I. Agartz, L. Westlye, U. Haukvik, A. Dale, et al., Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the enigma consortium, Molecular psychiatry 21 (4) (2016) 547–553.

2. R. Geevarghese, D. E. Lumsden, N. Hulse, M. Samuel, K. Ashkan, Subcortical structure volumes and correlation to clinical variables in parkinson's disease, Journal of Neuroimaging 25 (2) (2015) 275–280.

3. S. Goldman, L. M. OBrien, P. A. Filipek, I. Rapin, M. R. Herbert, Motor stereotypies and volumetric brain alterations in children with autistic disorder, Research in autism spectrum disorders 7 (1) (2013) 82–92.

4. X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, À. Rovira, Segmentation of multiple sclerosis lesions in brain mri: a review of automated approaches, Information Sciences 186 (1) (2012) 164–185.

5. D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, D. L. Collins, Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging, Medical image analysis 17 (1) (2013) 1–18.

6. P. Koolschijn, N. E. van Haren, G. J. Lensvelt-Mulders, H. Pol, E. Hilleke, R. S. Kahn, Brain volume abnormalities in major depressive disorder: A meta-analysis of magnetic resonance imaging studies, Human brain mapping 30 (11) (2009) 3719–3735.

7. M. Langen, H. G. Schnack, H. Nederveen, D. Bos, B. E. Lahuis, M. V. de Jonge, H. van Engeland, S. Durston, Changes in the developmental trajectories of striatum in autism, Biological psychiatry 66 (4) (2009) 327–333.

8. M. Deeley, A. Chen, R. Datteri, J. Noble, A. Cmelak, E. Donnelly, A. Malcolm, L. Moretti, J. Jaboin, K. Niermann, et al., Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study, Physics in medicine and biology 56 (14) (2011) 4557.

9. J. Dolz, L. Massoptier, M. Vermandel, Segmentation algorithms of subcortical brain structures on mri for radiotherapy and radiosurgery: a survey, IRBM 36 (4) (2015) 200–212.

10. J. M. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, D. Rueckert, A. D. N. Initiative, et al., Fast and robust multi-atlas segmentation of brain magnetic resonance images, Neuroimage 49 (3) (2010) 2352–2365.

11. H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, P. A. Yushkevich, Multi-atlas segmentation with joint label fusion, IEEE transactions on pattern analysis and machine intelligence 35 (3) (2013) 611–623.

12. K. O. Babalola, T. F. Cootes, C. J. Twining, V. Petrovic, C. Taylor, 3d brain segmentation using active appearance models and local regressors, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2008, pp. 401–408.

13. A. Rao, P. Aljabar, D. Rueckert, Hierarchical statistical shape analysis and prediction of sub-cortical brain structures, Medical image analysis 12 (1) (2008) 55–68.

14. J. Yang, J. S. Duncan, 3d image segmentation of deformable objects with joint shape-intensity prior models using level sets, Medical Image Analysis 8 (3) (2004) 285–294.

15. S. Powell, V. A. Magnotta, H. Johnson, V. K. Jammalamadaka, R. Pierson, N. C. Andreasen, Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures, Neuroimage 39 (1) (2008) 238–247.

16. J. Dolz, A. Laprie, S. Ken, H.-A. Leroy, N. Reyns, L. Massoptier, M. Vermandel, Supervised machine learning-based classification scheme to segment the brainstem on mri in multicenter brain tumor treatment context, International journal of computer assisted radiology and surgery 11 (1) (2016) 43–51.

17. M. Shakeri, S. Tsogkas, E. Ferrante, S. Lippe, S. Kadoury, N. Paragios, I. Kokkinos, Sub-cortical brain structure segmentation using f-cnn's, arXiv preprint arXiv:1602.02130.

18. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

19. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

20. D. Ciresan, A. Giusti, L. M. Gambardella, J. Schmidhuber, Deep neural networks segment neuronal membranes in electron microscopy images, in: Advances in neural information processing systems, 2012, pp. 2843–2851.

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

22. M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks, Medical Image Analysis.

23. S. Pereira, A. Pinto, V. Alves, C. A. Silva, Brain tumor segmentation using convolutional neural networks in mri images, IEEE transactions on medical imaging 35 (5) (2016) 1240–1251.

24. N. Lee, A. F. Laine, A. Klein, Towards a deep learning approach to brain parcellation, in: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE, 2011, pp. 321–324.

25. P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, I. Išgum, Automatic segmentation of mr brain images with a convolutional neural network, IEEE transactions on medical imaging 35 (5) (2016) 1252–1261.

26. F. Milletari, S.-A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Bötzel, et al., Hough-cnn: Deep learning for segmentation of deep brain regions in mri and ultrasound, arXiv preprint arXiv:1601.07014.

27. A. de Brebisson, G. Montana, Deep neural networks for anatomical brain segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 20–28.

28. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

29. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

30. T. Brosch, Y. Yoo, L. Y. Tang, D. K. Li, A. Traboulsee, R. Tam, Deep convolutional encoder networks for multiple sclerosis lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 3–11.

31. K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation, arXiv preprint arXiv:1603.05959.

32. D. Scherer, A. Müller, S. Behnke, Evaluation of pooling operations in convolutional architectures for object recognition, in: International Conference on Artificial Neural Networks, Springer, 2010, pp. 92–101.

33. T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: European conference on computer vision, Springer, 2004, pp. 25–36.

34. C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE transactions on pattern analysis and machine intelligence 35 (8) (2013) 1915–1929.

35. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062.

36. L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, A. L. Yuille, Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform, arXiv preprint arXiv:1511.03328.

37. B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 447–456.

38. S. M. Smith, Fast robust automated brain extraction, Human brain mapping 17 (3) (2002) 143–155.

39. J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the eighteenth international conference on machine learning, ICML, Vol. 1, 2001, pp. 282–289.

40. A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism, Molecular psychiatry 19 (6) (2014) 659–667.

41. B. Fischl, Freesurfer, Neuroimage 62 (2) (2012) 774–781.

42. E. H. Aylward, N. J. Minshew, K. Field, B. Sparks, N. Singh, Effects of age on brain volume and head circumference in autism, Neurology 59 (2) (2002) 175–183.

43. W. Sato, Y. Kubota, T. Kochiyama, S. Uono, S. Yoshimura, R. Sawada, M. Sakihama, M. Toichi, Increased putamen volume in adults with autism spectrum disorder, Frontiers in human neuroscience 8.

44. D. R. Hampson, G. J. Blatt, Autism spectrum disorders and neuropathology of the cerebellum, Frontiers in neuroscience 9.

45. R. Nicolson, T. J. DeVito, C. N. Vidal, Y. Sui, K. M. Hayashi, D. J. Drost, P. C. Williamson, N. Rajakumar, A. W. Toga, P. M. Thompson, Detection and mapping of hippocampal abnormalities in autism, Psychiatry Research: Neuroimaging 148 (1) (2006) 11–21.

46. C. M. Schumann, J. Hamstra, B. L. Goodlin-Jones, L. J. Lotspeich, H. Kwon, M. H. Buonocore, C. R. Lammers, A. L. Reiss, D. G. Amaral, The amygdala is enlarged in children but not adolescents with autism; the hippocampus is enlarged at all ages, The Journal of Neuroscience 24 (28) (2004) 6392–6401.

47. A. Hardan, N. Minshew, M. Keshavan, Corpus callosum size in autism, Neurology 55 (7) (2000) 1033–1036.

48. K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

49. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

50. J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, Theano: A cpu and gpu math compiler in python, in: Proc. 9th Python in Science Conf, 2010, pp. 1–7.

51. L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.

52. I. Wolf, M. Vetter, I. Wegner, T. Böttger, M. Nolden, M. Schöbinger, M. Hastenteufel, T. Kunert, H.-P. Meinzer, The medical imaging interaction toolkit, Medical image analysis 9 (6) (2005) 594–604.

53. A. R. Khan, N. Cherbuin, W. Wen, K. J. Anstey, P. Sachdev, M. F. Beg, Optimal weights for local multi-atlas fusion using supervised learning and dynamic information (superdyn): validation on hippocampus segmentation, NeuroImage 56 (1) (2011) 126–139.

54. A. R. Khan, L. Wang, M. F. Beg, Freesurfer-initiated fully-automated subcortical brain segmentation in mri using large deformation diffeomorphic metric mapping, Neuroimage 41 (3) (2008) 735–746.

55. Y. Huo, A. J. Plassard, A. Carass, S. M. Resnick, D. L. Pham, J. L. Prince, B. A. Landman, Consistent cortical reconstruction and multi-atlas brain segmentation, NeuroImage.

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

57. X. Han, B. Fischl, Atlas renormalization for improved brain mr image segmentation across scanner platforms, IEEE transactions on medical imaging 26 (4) (2007) 479–486.

58. M. G. Linguraru, T. Vercauteren, M. Reyes-Aguirre, M. Á. G. Ballester, N. Ayache, Segmentation propagation from deformable atlases for brain mapping and analysis, Brain Research Journal 1 (4) (2007) 1–18.

59. P.-L. Bazin, D. L. Pham, Homeomorphic brain image segmentation with topological and statistical atlases, Medical image analysis 12 (5) (2008) 616–625.

60. X. Artaechevarria, A. Munoz-Barrutia, C. Ortiz-de Solórzano, Combination strategies in multi-atlas image segmentation: application to brain mr data, IEEE transactions on medical imaging 28 (8) (2009) 1266–1277.

61. C. Ciofolo, C. Barillot, Atlas-based segmentation of 3d cerebral structures with competitive level sets and fuzzy control, Medical image analysis 13 (3) (2009) 456–470.

62. M. R. Sabuncu, B. T. Yeo, K. Van Leemput, B. Fischl, P. Golland, A generative model for image segmentation based on label fusion, IEEE transactions on medical imaging 29 (10) (2010) 1714–1729.

63. B. Patenaude, S. M. Smith, D. N. Kennedy, M. Jenkinson, A bayesian model of shape and appearance for subcortical brain segmentation, Neuroimage 56 (3) (2011) 907–922.

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

65. A. J. Asman, F. W. Bryan, S. A. Smith, D. S. Reich, B. A. Landman, Groupwise multi-atlas segmentation of the spinal cords internal structure, Medical image analysis 18 (3) (2014) 460–471.

66. J. Wang, C. Vachet, A. Rumple, S. Gouttard, C. Ouziel, E. Perrot, G. Du, X. Huang, G. Gerig, M. A. Styner, Multi-atlas segmentation of subcortical brain structures via the autoseg software pipeline, Frontiers in neuroinformatics 8 (2014) 7.

67. S. Bao, A. C. Chung, Multi-scale structured cnn with label consistency for brain mr image segmentation, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (2016) 1–5.